RICE UNIVERSITY

Towards Adaptive Resolution Modeling of Biomolecular Systems in their Environment

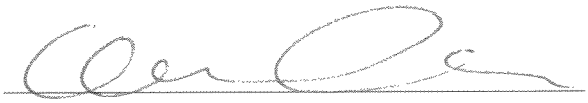
by

Bradley Paul Lambeth, Jr.

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Approved, Thesis Committee:



Dr. Cecilia Clementi, Chair
Professor of Chemistry and Chemical & Biomolecular Engineering

Dr. Matteo Pasquali
Professor of Chemical & Biomolecular Engineering and Chemistry

Dr. Gustavo Scuseria
Robert A Welch Professor of Chemistry and Physics

Houston, Texas

May, 2012

ABSTRACT

Towards Adaptive Resolution Modeling of Biomolecular Systems in their Environment

by

Bradley Paul Lambeth, Jr.

Water plays a critical role in the function and structure of biological systems. Current techniques to study biologically relevant events that span many length and time scales are limited by the prohibitive computational cost of including accurate effects from the aqueous environment. The aim of this work is to expand the reach of current molecular dynamics techniques by reducing the computational cost for achieving an accurate description of water and its effects on biomolecular systems.

This work builds from the assumption that the "local" effect of water (e.g. the local orientational preferences and hydrogen bonding) can be effectively modelled considering only the atomistic detail in a very limited region. A recent adaptive resolution simulation technique (AdResS) has been developed to practically apply this idea; in this work it will be extended to systems of simple hydrophobic solutes to

determine a characteristic length for which thermodynamic, structural, and dynamic properties are preserved near the solute. This characteristic length can then be used for simulation of biomolecular systems, specifically those involving protein dynamics in water. Before this can be done, current coarse grain models must be adapted to couple with a coarse grain model of water.

This thesis is organized in to five chapters. The first will give an overview of water, and the current methodologies used to simulate water in biological systems. The second chapter will describe the AdResS technique and its application to simple test systems. The third chapter will show that this method can be used to accurately describe hydrophobic solutes in water. The fourth chapter describes the use of coarse grain models as a starting point for targeted search with all-atom models. The final chapter will describe attempts to couple a coarse grain model of a protein with a single-site model for water, and it's implications for future multi-resolution studies.

Acknowledgments

I would like to thank my wife, Christine, for supporting me throughout my graduate career. Her persistent work ethic was a continual encouragement for me as I struggled with scientific puzzles which could not be solved in the matter of hours we had become accustomed to during our undergraduate years. I dedicate this thesis to you.

My advisor, Cecilia, provided me with so many learning opportunities in the form of well-designed projects, collaboration contacts, invitations to conferences, and extended studies abroad. The long trips to Germany and Los Angeles were particularly productive times for each of the projects described in this work.

I am greatly indebted to Paul Ledbetter for his extensive help with much of the algorithms and theory presented in this work. Any time I needed to write a new piece of code, chances are he had written something similar already and could provide guidance on what to try first. It was a pleasure to work with him on his PYP project.

Finally, I would like to thank my collaborators Christoph Junghans, Luigi Delle Site, and Kurt Kremer, as well as my fellow group members Wenwei Zheng and Mary Rohrdanz. Many of the ideas that resulted in breakthroughs came to fruition after explaining to each of them what was not working.

Contents

Illustrations

Tables

Chapter 1

Water in Molecular Dynamics

1.1 Introduction

Water has a long history of research, especially in the last few decades as techniques for studying biomolecules have become increasingly robust. This simple liquid is capable of forming complex hydrogen bond networks which directly affect the way in which biological molecules move and function [1, 2]. For small solutes, the hydrogen bond network is capable of simply encapsulating the solute, but for larger solutes such as proteins the network must deform to include the solute [3]. Many of the theories of protein dynamics are based on the idea of a "hydrophobic effect" which suggests that a protein orders itself in such a way as to bury as many hydrophobic residues as possible, while exposing hydrophilic residues capable of forming hydrogen bonds with the environment's hydrogen bond network [4]. The details of the disruption and formation of hydrogen bonds are vitally important to the understanding of protein dynamics.

Ideally one would include as much information about the biological environment as possible, such as ion concentrations and pH balance, but the computational barrier of including water alone has yet to be surpassed for biologically relevant time scales, and

by including so much detail, it is also difficult to understand which effects are most important to the dynamics. This chapter will begin by describing the state of the art techniques for including water in molecular dynamics simulations, focusing on the use of all-atom models to achieve structural details such as water-mediated interactions and hydrogen bond networks. The chapter will conclude with a description of various techniques that have been recently proposed for reducing the level of detail in such a way to preserve most of the relevant properties for water.

1.2 Explicit Water Models

It is important to note that no water model can reproduce all of the physical properties of water; in fact none of the models are capable of matching the experimental radial distribution function and internal energy values at the same time [5]. Development of explicit water models has been ongoing since Bernal and Fowler proposed the first model in 1933 [6]. Rahman and Stillinger proposed the ST4 model in their work on examining the hydrogen bond networks of water [7]. The ST4 model contains four charge sites - one for each hydrogen and lone pair of electrons - in a tetrahedral arrangement to maximize hydrogen bonds. The simple point charge (SPC) model was introduced by Berendsen and co-workers in 1981 and an important work comparing several other models including TIP3P and TIP4P was done by Jorgensen in 1983 [8]. These three models, along with the extended SPC (SPC/E) model [9], have experienced the majority of use in molecular dynamics simulations [10].

The TIP3P and TIP4P models were primarily optimized to reproduce the heat of vaporization of water at standard conditions, with the only difference being the movement of the negative partial charge from the oxygen to a virtual site on the bisector of the HOH angle for TIP4P. This modification increases the number of distances that are required to calculate water-water interactions to 10 from 9 but achieves a better fit for the experimental RDF. The SPC/E model modifies the point charges of SPC slightly to account for the averaged polarization effects, resulting in better agreement with the experimental RDF at the second peak. Recently the TIPnP models have been further modified with an additional lone pair virtual site which mimics the early tetrahedral models. With this additional parameter the TIP5P model [11] reproduces the density anomaly of water at 277K, with slight improvements in the RDF as well. However, each additional parameter increases the computational cost of the model.

The computational cost of a water model is directly related to the number of distances that must be calculated to evaluate the effective pair potential. The TIP3P and SPC/E pair potentials require the least number of distances for water-water interactions

$$V_{\alpha\beta}(r_{ij}, r_{OO}) = \sum_{i\in\alpha}\sum_{j\in\beta}\frac{q_i q_j e^2}{4\pi\varepsilon_0 r_{ij}} + \frac{C_{12}}{r_{OO}^{12}} - \frac{C_6}{r_{OO}^{6}} \tag{1.1}$$

where the first term is the electrostatic contribution with q_i and q_j as constant point charges of atoms i and j, r_{ij} is the distance between atoms i and j, ε_0 is the permittivity of free space, and C_{12} and C_6 are the repulsive and dispersion parts of

Table 1.1 : Parameters of explicit water models

Parameter	TIP3P	TIP4P	TIP5P	SPC	SPC/E
r_{OH} (Å)	0.9572	0.9572	0.9572	1.0	1.0
∠HOH (°)	104.52	104.52	104.52	109.47	109.47
$10^3 C_6$ (kJ/mol nm^6)	2.4889	2.5499	2.4699	2.6171	2.6171
$10^6 C_{12}$ (kJ/mol nm^{12})	2.4352	2.5084	2.2783	2.6331	2.6331
q_{Oxygen} (e)	-0.834	-	-	-0.820	-0.8476
$q_{Hydrogen}$ (e)	+0.417	+0.520	+0.241	+0.410	+0.4238
$q_{Virtual}$ (e)	-	-1.04	-0.241	-	-

the Lennard-Jones interaction between oxygen atoms. Table 1.1 summarizes the parameters of the most popular water models. Extensions of these models can include internal degrees of freedom such as bond and angle stretching and are often referred to as "flexible" models.

1.3 Implicit Solvation

Implicit solvation methods completely remove the atomic detail of the solvent and replace it with a field that implicitly reproduces the same free energy of a solvated molecule. The basic idea of implicit models is that ΔG_{solv} can be approximated as

$$\Delta G_{solv} = \Delta G_{el} + \Delta G_{nonel} \tag{1.2}$$

where ΔG_{el} is the free energy associated to solvating the molecule and ΔG_{nonel} is the free energy of the molecule in vacuum.

A number of factors enter the approximation of this solvating term, ΔG_{el}; one of the most important is captured by the exposed surface area. The Generalized Born (GB) model is a popular implicit model that takes in to account the exposed surface area effect by treating each atom in a molecule as a sphere of radius R_i with a charge q_i at its center. The interior of the atom is assumed to be a material of uniform dielectric constant 1, while the molecule is assumed to be surrounded by a solvent of a high dielectric constant ε. The GB approximation assumes the free energy of solvation is of the form

$$\Delta G_{el} = -\frac{1}{2} \sum_{i,j} \frac{q_i q_j}{f_{GB}(r_{ij}, R_i, R_j)} \left(1 - \frac{e^{-\kappa f_{GB}}}{\varepsilon}\right) \tag{1.3}$$

where q_i and q_j are the charges of the respective atoms, r_{ij} is the distance between the atoms, R_i and R_j are the respective Born radii, κ is the Debye-Huckel screening parameter, and f_{GB} is a smoothing function which depends on the various Generalized Born models [12]. An application of the Poisson Boltzmann equation to achieve a better approximation of the nature of the solvent can lead to more accurate results, but the computational cost of solving the nonlinear system of equations at each step reduces the application of the method considerably [13].

Regardless of the level of detail of the approximation, the implicit solvation methods by construct remove the atomic detail of the solvent at the solvent interface. The detailed interactions of water near the solute interface, and the resulting hydrogen

bond network, cannot be captured, thus limiting the application of the models to systems where local structure of water is not important. As the goal of this project is to develop a method that allows to capture the local structure of water, implicit water methods will not be considered, at least at the local scale.

1.4 Effective Interactions

Various coarse grain methods for biomolecular systems reduce the computational cost associated with the model of atomic detail by introducing pseudoatoms which effectively describe the collective effect of groups of atoms. The definition of the explicit functional form for the effective interactions between pseudoatoms is the main challenge in a coarse grain model, but it is usually determined by comparing with experimental data in such a way that native structure and other thermodynamic properties are reproduced at least approximately. As in the work of Matysiak and Clementi [14], because the experimentally determined target structure depends on the environmental effects, the local effect of water is believed to be included in the effective interactions, thus negating the need for explicit water. However, these methods are similar to the implicit solvation methods in that they do not capture the structure of water at the solute interface. Explicit models must be used to capture this detail.

1.5 Coarse Graining Water

Because of the high computational cost of calculating the very long range electrostatic contribution to the all-atom potential energy, efforts have been made to coarse grain these models by introducing "pseudoatoms" that group together several atoms in the molecule (or whole molecule) while reproducing as much of the physical properties as possible. A few of the methods that are currently used to achieve effective potentials for coarse grain models are described below.

1.5.1 Force Matching

The goal of the Force Matching procedure introduced by Voth [15] is to create an effective potential which matches the combined force of all the atoms in the all-atom model on molecules. With additional constraints the procedure can also reproduce the pressure of the system [16]. The primary shortcoming of this method is that it does not include any requirement to reproduce any specific physical property, and as a result it often fails to do so . For example, when compared to all-atom models, the RDF and average tetrahedral packing parameter (q) obtained with force matching has a very poor fit for TIP3P water [17]. For these reasons this method will not be used for the present study of water.

1.5.2 Iterative Boltzmann Inversion

The Iterative Boltzmann Inversion (IBI) method is capable of generating a potential that reproduces the center of mass radial distribution function (RDF_{cm}) of the all-atom model at a given state point [18]. The basic idea is that in the ideal gas limit, the effective potential for molecules is the potential of mean force. Simulations are performed for the coarse grained model starting with the potential of mean force as a first approximation for the non-ideal gas all-atom system

$$V_0(r) = -k_B T \ln g^{tgt}(r) \tag{1.4}$$

where $g^{tgt}(r)$ is the target radial distribution function of the all-atom system. The effective potential is updated after each step by using the correction

$$V_{i+1}(r) = V_i(r) + k_B T \ln \left(\frac{g_i(r)}{g^{tgt}(r)} \right) \tag{1.5}$$

.

Additionally, a weak constant force is added at each step to match the pressure of the all-atom system. The method converges when the update to the potential is negligible, which corresponds to an exact match of the RDF_{cm}. The IBI procedure is limited in the fact that, by definition, the RDF is the only property that is expected to be reproduced, while structural or thermodynamic properties may not be correctly modeled [19].

1.5.3 Inverse Statistical Mechanics

The inverse statistical mechanics method was originally proposed by Lyubartsev and Laaksonen [20], and it is similar in nature to the IBI procedure in that the radial distribution function is the primary target of the optimization procedure. However, the definition of the method is much more general and can be easily modified to take in to account other properties. Because of this increased flexibility, the inverse statistical mechanics method is most appropriate for coarse grain simulations of water in which the optimization of many parameters is necessary, as it is in this work.

Given a initial guess Hamiltonian H which has parameters ϵ, if it can be assumed that the Hamiltonian is not far from the expected result, a generic thermodynamic average X can be optimized at each step by a first order correction of the form

$$\sum_{\alpha}(X_{\alpha}^{tgt}-X_{\alpha}^{i}) = -\beta\sum_{i=1}^{n}\left(\left\langle X_{\alpha}\frac{\delta H(\epsilon_1,...,\epsilon_n)}{\delta\epsilon_i}\right\rangle - \langle X_{\alpha}\rangle\left\langle\frac{\delta H\left(\epsilon_1,...,\epsilon_n\right)}{\delta\epsilon_i}\right\rangle\right)\Delta\epsilon + O\left(\epsilon^2\right) \tag{1.6}$$

where $\beta = \frac{1}{k_B T}$. For the simple case of an effective potential targeted at reproducing the radial distribution function, one could use a potential of the form

$$H = \sum_{\alpha} K_{\alpha} S_{\alpha} \tag{1.7}$$

where K_{α} are the parameters of the Hamiltonian and S_{α} is the pair density derived from the radial distribution function as

$$\langle S_{\alpha}\rangle = \frac{4\pi r^2 g(r) N^2}{2V} \tag{1.8}$$

In this simple case the first order correction to the potential simplifies to

$$\Delta S_\alpha = \sum_\gamma -\beta \left(\langle S_\alpha S_\gamma \rangle - \langle S_\alpha \rangle \langle S_\gamma \rangle \right) \Delta H_\gamma \tag{1.9}$$

To optimize other physical properties, the form of the Hamiltonian must be changed. The effective potential for TIP3P water obtained using this procedure has been shown to better reproduce the tetrahedral packing parameter q[21], described in Section 2.3.3, than the other methods, and will thus be used in these studies.

1.6 Summary

A brief account has been given of the inclusion of water in molecular dynamics simulations and the goals of such endeavors. As mentioned above, no water model exists that can reproduce all physical properties for water. However, explicit models are the only available method to capture the local structural details of water at wet/dry interfaces, and because of the relevance of such details to biological phenomena these methods will be used instead of implicit solvation methods. In order to reduce the computational cost of including all-atom detail, efforts have been made to coarse grain away the unimportant degrees of freedom for bulk water. These models, which cannot reproduce all of the structural details of the more complex models, are however much less expensive. The remainder of this thesis will describe the efforts taken to combine the benefits of inexpensive one-site simulations with structurally important all-atom systems by coupling two resolutions with either adaptive resolution techniques or systematic comparison.

1.7 Key achievements of this thesis

In Chapter 3, an application of the AdResS methodology for hydrophobic solutes is presented for the first time. The positive results from this study show that the AdResS method is capable of improving the efficiency for realistic systems while maintaining the important information without significant degradation. In order to accomplish these simulations, the AdResS algorithm was integrated in to one of the most efficient molecular dynamics packages currently available, allowing easy access to the method for the scientific community.

In Chapter 4, a systematic comparison of a coarse grain model with it's all-atom counterpart shows the efficacy of combining these resolutions to explore the entire free energy landscape. The computational cost of the all-atom model is overcome by reconstructing coarse grain structures at important regions discovered during the coarse grain simulations. The computational cost could be further reduced by modeling the system in a hybrid fashion with both all-atom and coarse grain models present in the same simulation. This requires an coarse grain model capable of coupling to a coarse grain model of water, and a hybrid model for the protein.

In Chapter 5, a coarse grain model of a protein is coupled for the first time to a single-site model for water. The resulting model reveals information about the transition state which is not easily grasped with vacuum coarse grain models. The reduced computational cost of the system allows for microsecond exploration of the free energy landscape, encouraging the addition of a hybrid model for the protein to

allow high resolution investigation in functionally relevant regions, such as binding sites of chromophores.

Chapter 2

Adaptive Resolution Simulations

2.1 Introduction

Many phenomena in biology, chemistry, and materials science involve processes occurring on atomistic length and time scales that affect structural and dynamical properties on mesoscopic scales extending far beyond atomistic ones. Computer simulations of very large systems with full atomistic detail is often either infeasible, undesirable, or unnecessary. This has lead to the development of coarse grained (CG) models which allow for the study of mesoscale changes at large length scales. The reduced computational cost of CG models allows full system equilibration over mesoscopic time scales. The goal for these reduced models is to reproduce important physicochemical features of a selected system by trading a large gain in simulation speed with the introduction of some approximations.

Recently CG models have been used in combination with models of higher resolution in so-called "multi-scale" methods in order to exploit their complementary benefits. The linking of high and low resolution models requires that they be structurally consistent. The theoretical methods that are employed to study such multiscale systems span from quantum-mechanical to macroscopic statistical approaches.

A common limitation of these methods is that the regions of the system treated at different resolution levels are often fixed and do not allow for particle exchange. Recently, however, a particle-based adaptive resolution simulation scheme (AdResS) was proposed for hybrid all-atom/CG molecular dynamics simulations which allows for particle exchange between the different resolution regions.

2.2 AdResS: Theoretical Background

Consider a system of N molecules in a volume V, modeled at an atomistic level, in which full atomistic detail is not needed in a certain subvolume V_A. Thus the number of degrees of freedom (DOF) n per molecule is lower in the CG region A and higher in the all-atom region B. For simplicity, one can divide the simulation volume V into two equally large slabs A and B. In order to simulate the two different resolutions for the same system, thermodynamic equilibrium between the two regions must be maintained to avoid spurious effects.

For constant molecular density the chemical potentials must be equal in both regions, which guarantees that the molecules experience no spurious driving force into one region and keeps the system homogeneous across the box. The artificial resolution boundary must be essentially invisible, in that the molecules have to cross the border between the two regions without experiencing any barrier. To achieve this, conditions analogous to a two-phase coexistence, μ_A=μ_B, p_A=p_B , T_A=T_B, must be satisfied, where, μ_A , p_A , and T_A and μ_B , p_B , and T_B are the chemical potentials, pressures,

and temperatures of the liquid in the CG and atomistic domains, respectively.

To smoothly couple the different regions, a two-stage procedure is applied [22]. First, CG molecules are defined and an effective pair potential $U_{cg}(r)$ between the CG molecules is derived such that the statistical properties (i.e., the density, pressure, and temperature) of the corresponding all-atom system are accurately reproduced. Depending on the system, other properties such as the radial distribution functions should also match as closely as possible. Different methods for deriving these effective potentials have been presented in the previous chapter (Section 1.5) for the case of water.

Second, an interface layer is introduced between the atomistic and CG region which contains hybrid molecules. These hybrid molecules correspond to an all-atom molecule with an additional massless center of mass particle serving as an interaction site. The transition is then governed by a weighting function $w(x) \in [0, 1]$ which interpolates the molecular interaction forces between the two regions and assigns a resolution "identity" to the molecule. The weighting function is defined in such a way that high resolution molecules have $w=1$, low resolution have $w=0$, and values $0<w<1$ are assigned in the transition region. Interpolating the forces acting between the center of mass interaction sites for molecules α and β leads to the definition of a hybrid force

$$F_{\alpha\beta} = w(x_\alpha)w(x_\beta)F^{ex}_{\alpha\beta} + [1 - w(x_\alpha)w(x_\beta)]\, F^{cg}_{\alpha\beta} \tag{2.1}$$

where $F_{\alpha\beta}$ is the total intermolecular force acting between the center of mass of the

molecules α and β, $F^{ex}_{\alpha\beta}$ is the sum of all pair atom interactions between explicit atoms of the molecule α and explicit atoms of the molecule β, $F^{cg}_{\alpha\beta}= -\nabla U^{cg}_{\alpha\beta}$ is the effective pair force between the two molecules, and x_α and x_β are the positions of the center of mass sites of the molecules α and α, respectively. By interpolating the force in this way, AdResS by construction satisfies Newton's third law, which is crucial for the proper diffusion of molecules across the resolution boundaries.

Because the total pair force depends not only on the relative distances but also on the positions of the molecules relative to some reference, it is not conservative, and it is not possible to associate a corresponding potential as the work done by this force depends on the physical path taken by the molecule in the transition region. However, the grand-canonical potential $\Omega=-\nabla pV$ is still a well-defined and conserved quantity in the AdResS approach and can be used to satisfy two-phase coexistence requirements, eliminating any spurious flux at the boundary between the atomistic and CG regions. A molecule gains or loses, depending on whether it leaves or enters the CG region, its equilibrated rotational and vibrational DOFs when it crosses a boundary between the different regions, while retaining its linear momentum. Because of the change in DOFs, this change in resolution requires the addition or removal of latent heat and thus must be accompanied with a thermostat that couples locally to the particle motion.

One could also attempt to introduce a mixing scheme with a similar interpolation technique but working with the interaction potentials rather than the forces. This in-

deed has been recently attempted [23], however such an approach leads to a violation of Newton's third law, and consequently to a non-conservation of the linear momentum [24]. Furthermore, in that case the total force would not only be a function of the gradients of the explicit and CG potentials, but would also depend linearly on the respective potentials themselves. The idea of decoupling DOFs rather than switching them on and off is attractive at first sight, but the concept only works for DOFs that are eigenmodes of the molecules without any coupling to the surrounding [25]. Because of these restrictions, the force interpolation scheme was chosen for AdResS.

2.3 Example Applications

2.3.1 Tetrahedral Molecule Liquids

The first proof of principle to test the AdResS idea used a liquid of simple model tetrahedral molecules [26]. The model liquid mimics a typical liquid such as methane. Each molecule in this system comprises four identical atoms with mass m_0. The intramolecular interactions are regulated by the repulsive Weeks-Chandler-Andersen potential

$$U_{rep}^{ex}(r_{i\alpha j\beta}) = \begin{cases} 4\varepsilon \left[\left(\frac{\sigma}{r_{i\alpha j\beta}} \right)^{12} - \left(\frac{\sigma}{r_{i\alpha j\beta}} \right)^{6} + \frac{1}{4} \right] & ; r_{i\alpha j\beta} \leq 2^{1/6}\sigma \\ 0 & ; r_{i\alpha j\beta} > 2^{1/6}\sigma \end{cases} \tag{2.2}$$

where σ and ϵ are the standard Lennard-Jones parameters of length and energy, and $r_{i\alpha j\beta}$ is the distance between the atom i of the molecule α and the atom j of the molecule β. The atoms are connected by finite extensible nonlinear elastic (FENE)

bonds

$$U_{bond}^{ex}(r_{i\alpha j\alpha}) = \begin{cases} -\frac{1}{2}kR_0^2 \ln\left[1 - \left(\frac{r_{i\alpha j\alpha}}{R_0}\right)\right] & ; r_{i\alpha j\alpha} \leq R_0 \\ 0 & ; r_{i\alpha j\alpha} > R_0 \end{cases} \tag{2.3}$$

with R_0=1.5σ and k=30$\frac{\varepsilon}{\sigma^2}$.

Medium density and high density liquids with molecular number densities of ρ_m=$\frac{0.1}{\sigma^3}$ and ρ_h=$\frac{0.175}{\sigma^3}$, respectively, were investigated. Besides the linear slab-like geometry, a spherical cavity-like geometry for the system was also tested. The latter example is useful for studies of macromolecules embedded in a solvent, whereas the former can be readily extended to studies of liquids near surfaces.

An explicit (EX) atomistic model was mapped to a CG model comprised of N one-particle molecules. A CG molecule α in the system has a mass M_α=4m_0 equal to the total mass of an explicit tetrahedral molecule. The number of non-bonded interactions is significantly reduced, and all the rotational and vibrational DOFs of the explicitly resolved tetrahedral molecules are removed. The effective CG potential $U_{cg}(r)$ is derived by using the IBI technique (see Section 1.5.2) in such a way that the center of mass radial distribution function (RDF_{cm}) and pressure of the CG system match the corresponding RDF_{cm} and pressure of the explicit system at a given density and temperature. This means that at any chosen point, the equations of state for the EX and CG systems coincide.

The atomistic and mesoscopic regions are then coupled using the AdResS method. Figure 2.1a shows the EX, hybrid, and CG representation of the tetrahedral molecule.

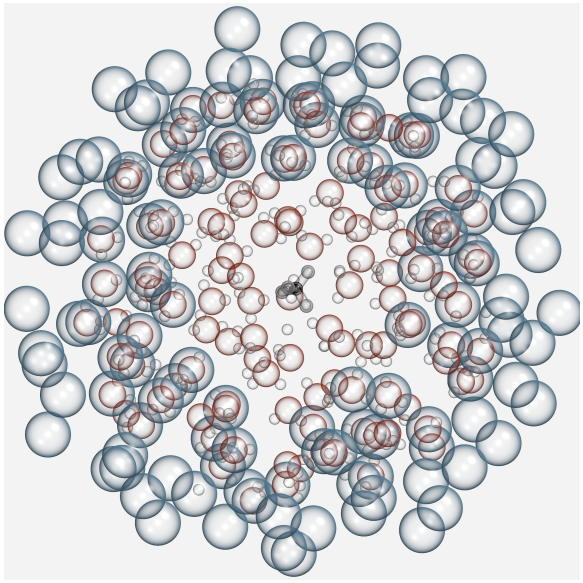


Figure 2.1 : Example AdResS Setup for solvated methane. Methane is resolved in all-atom detail, as well as the first few layers of water. A hybrid layer separates the all-atom region from the single-site coarse grain water in the bulk.

Figure 2.1 shows the simulation box in a spherical cavity-like geometry. An appropriate weighting function - *w(x)* for the slab-like geometry, *w(r)* for the spherical cavity-like geometry - is defined which smoothly interpolates between the regions and has zero slope at the boundaries of the transition region. Within the transition region, the pair force between the atoms $i\alpha$ and $j\beta$ of either explicit or hybrid molecules α and β, respectively, and the massless center of the molecule, is interpolated as described earlier.

The density profile throughout the transition region must be kept constant to avoid any kinetic barrier for a free exchange of particles. The linear combination of the explicit and coarse grain potential creates a slight dependence on the value of w for the pressure. The pressure increase is most prominent for $w = \frac{1}{2}$, so an interface

pressure correction is derived such that the pressure for a system of hybrid molecules with identity $w = \frac{1}{2}$ exactly reproduces the correct pressure and RDF_{cm}.

This sample test showed that AdResS can in principle be used to model homogeneous liquid systems. This test did not show, however, how a solute can be embedded in the system, or whether the method can be applied to more complex systems such as liquid water. The work described in Section 2.3.3 extended this approach to a realistic representation of water.

2.3.2 Solvated Polymer

The AdResS approach can be generalized to the study of a polymer chain in solution [27]. The chain is surrounded by solvent with atomistic resolution. As the polymer moves, the sphere of atomistic resolution moves together with the center of mass of the chain. The chain is free to move around, and the explicit resolution sphere is much smaller than the overall simulation volume. Only the solvent in the vicinity of the polymer is always represented with a sufficiently high level of detail so that the local interactions between a solvent and a solute are correctly taken into account, whereas solvent further away from the polymer, at which the high resolution is no longer required, is represented on a more CG level. The macromolecule is represented by a generic flexible polymer chain embedded in a solvent, for instance the system of tetrahedral molecules which was introduced in the previous section. To account for the movement of the high-resolution scheme with the polymer, the AdResS force

interpolation is generalized to:

$$F_{\alpha\beta} = w\left(|R_\alpha - R|\right) w\left(|R_\beta - R|\right) F^{ex}_{\alpha\beta} + \left[1 - w\left(|R_\alpha - R|\right) w\left(|R_\beta - R|\right)\right] F^{cg}_{\alpha\beta} \quad (2.4)$$

where R_α, R_β, and R are the center of mass positions of the molecules α,β, and the whole polymer, respectively. The size of the explicit region is set such that the polymer is always surrounded by explicit solvent, so that the interpolation only applies to solvent molecules.

Excellent agreement was found between the scattering function *S(q)* obtained with the AdResS and explicit simulations, and the mean square displacements of the polymer beads was found to be close to the expected power law of 0.6, showing that the conformational statistics of the embedded polymer in solution were properly reproduced using the AdResS method. The solvent used to test this extension of AdResS was still a hypothetical liquid though; the next example shows that the scheme can also be used for more complex liquids, like water.

2.3.3 Liquid Water

The AdResS approach can be extended to study polar liquids with long-range electrostatic interactions, i.e. TIP3P water [28]. In order to localize the long-range electrostatic interactions, the reaction field method needs to be used, in which all molecules with the charge center outside a spherical cavity of a molecular-based cut-off radius of R_c (typical values for $R_c = 9$Å) are treated as a dielectric continuum with a dielectric constant ε_{rf} [29, 30]. The electrostatic force acting on a charge $e_{i\alpha}$,

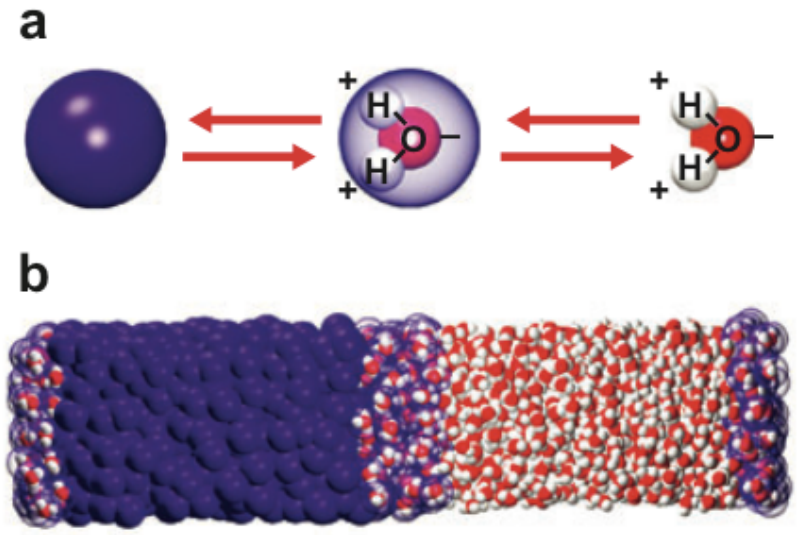


Figure 2.2 : AdResS representation of TIP3P water[28]

at the center of the cutoff sphere, due to a charge $e_{j\beta}$ within the cavity is

$$F^{ex}_{C_{i\alpha j\beta}}(r_{i\alpha j\beta}) = \frac{e_{i\alpha}e_{j\beta}}{4\pi\varepsilon_0}\left[\frac{1}{r^3_{i\alpha j\beta}} - \frac{1}{R_c^3}\frac{2(\varepsilon_{rf}-1)}{1+2\varepsilon_{rf}}\right] r_{i\alpha j\beta} \tag{2.5}$$

In order to use the AdResS procedure, a very realistic CG model needs to be used to allow a very fast and smooth equilibration of the atomistic DOFs adaptively added in the hybrid region. For this purpose a new single-site water model was developed that reproduces remarkably well the essential thermodynamic and structural features of water, as obtained by detailed all-atom simulations using the rigid TIP3P water model. Figure 2.2a describes the AdResS representation of TIP3P and Figure 2.2b describes the slab-like geometry of the multi-scale simulation.

To derive the effective potential between CG molecules, an iterative inverse statistical mechanics approach described earlier (Section 1.5.3) was used. Perfect agreement

between the all-atom and CG RDF_{cm}s was reached using the optimized effective potential. The effective potential has the primary minimum at approximately 2.8Å, corresponding to the first peak in the RDF_{cm}. The slightly weaker and significantly broader minimum at 4.5Å corresponds to the second hydration shell. As discussed earlier, different procedures for obtaining the effective potential can produce different results. The combined effect of the two minima leads to a local packing close to that of the all-atom TIP3P water. To date, this CG water model obtained with the IMC method is the one which best reproduces water properties at standard conditions [21].

To more thoroughly quantify the structural properties of the model which are not completely defined by the RDF, the distribution of the angle θ between the center of mass of three nearest-neighbor molecules and the orientational order parameter q were computed. The parameter q measures the extent to which a molecule and its four nearest neighbors adopt a tetrahedral arrangement, a feature which is essential to the physico-chemical properties of water.

The agreement for both structural parameters (RDF and q) was better than previous effective potentials for one-site water and could not be improved significantly by using more complex CG models (two-site, three-site), thus justifying the choice of the one-site model. For the hybrid simulation, the EX region had values of $\langle|\theta|\rangle_{ex} = 109.7 \pm 24.6$ and $\langle|q|\rangle_{ex} = 0.54 \pm 0.2$, and the CG region had values of $\langle|\theta|\rangle_{cg} = 103.4 \pm 30.7$ and $\langle|q|\rangle_{cg} = 0.42 \pm 0.2$. The reference bulk values for fully atomistic TIP3P simulations are $\langle|\theta|\rangle_{bulk} = 109.4 \pm 24.6$ and $\langle|q|\rangle_{bulk} = 0.55 \pm 0.2$. Most

importantly, the CG water did not adversely influence the structure in the all-atom region. The normalized density for the hybrid system was found to be homogeneous in the CG and EX regions with very small oscillations in the transition region, similar to the previous test cases described earlier.

The time evolution of a diffusion profile for molecules that were initially localized at the interface layer was computed to demonstrate the free exchange of molecules between the different regions. It was shown that these molecules spread out asymmetrically with time, arising from the difference in diffusion coefficient between the EX and CG regions due to the difference in DOFs. The diffusion could be matched with a position dependent Langevin thermostat, which effectively slows down the coarse grain molecules; however, when diffusion properties are not the focus of the study this difference in time scale can be advantageous for reaching longer effective simulation times in systems in which multiple length and time scales are intrinsically present. Along these lines AdResS can be applied to any other flexible or rigid nonpolarizable classical water model (e.g., SPC, SPC/E, TIP4P).

2.4 Summary

The AdResS method has been shown to accurately reproduce essential thermodynamic properties for simulations of homogeneous systems and one simple application of a polymer in solvent. The basic steps of AdResS involve a parametrization of an effective pair potential for the molecules which are to be coarse grained, and a defini-

tion of a simulation environment in which the all-atom region exchanges with and is supported by a bulk CG volume, separated by a hybrid transition region. There are several questions remaining to be answered, however, before such a method can be applied to biomolecular systems. For instance, what are the optimal values for the size of the all-atom and hybrid regions? How can multiple adaptive regions interact and mix? What are the effects of the approximation on the electrostatics when a macromolecule distorts the reaction field assumption? The following chapter will describe the first application of AdResS to a system with hydrophobic and hydrophilic molecules.

Chapter 3

Locality of Hydrogen Bond Networks at Hydrophobic Interfaces

3.1 Introduction

Liquid water is capable of forming highly complex hydrogen bond networks which directly affect the way in which biological molecules move and function. [1, 2] In this context understanding the solvation of hydrophobic molecules is a key to understand crucial processes occurring in (bio-)molecular systems. [31, 32, 33] In general, the molecular structure of liquid water around a solute molecule results from a competition between the disruption of the local tetrahedral order of bulk water and the formation of a two dimensional surface-like order at the solute interface. For very small hydrophobic solutes, such as methane, water molecules can encapsulate the guest molecule. [34] For increasingly larger solutes, the structure of water close to the solute eventually approaches the limiting case of an infinite hydrophobic surface. [35, 3, 36]

An important question related to the structure of water around a hydrophobic solute concerns the locality of the hydrogen bond network, that is, whether or not the solvation structure is determined by the surrounding bulk. In order to address

this computationally a tool is needed that can slowly switch on and off the hydrogen bonds in a well defined region around the solute, so that their relevance on the rest of the network can be determined in an unequivocal way. Switching on and off the hydrogen bond degrees of freedom must occur without affecting the thermodynamic equilibrium of the whole system and, if switched off only in the bulk, thermodynamic equilibrium must be assured between the regions of different resolution. Such a tool is provided by the adaptive resolution scheme (AdResS) [22], that allows to interface regions with different molecular representations (e.g. atomistic and coarse-grained) while maintaining free exchange of particles and equilibrium between the regimes, that is, the two different regions maintain equal temperature, pressure, and density. [37]

For the problem treated here AdResS is used to interface an atomistic model of water which explicitly forms hydrogen bonds with a coarse-grained (CG) spherical representation of water, which does not have any directional interactions. This allows to systematically determine the role of hydrogen bonds of the bulk onto the structure of water around hydrophobic solutes of different sizes. In particular we investigate the role of two common water-solute interactions, a Lennard-Jones (LJ) as well as a short ranged purely repulsive potential. For both interaction types the structure of the water layer close to the solute has to accommodate the geometrical constraints induced by the solvated molecule *.

*This chapter is adapted from Lambeth et al. [38], and I greatly acknowledge my co-authors C. Junghans, K. Kremer, C. Clementi, and L. Delle Site.

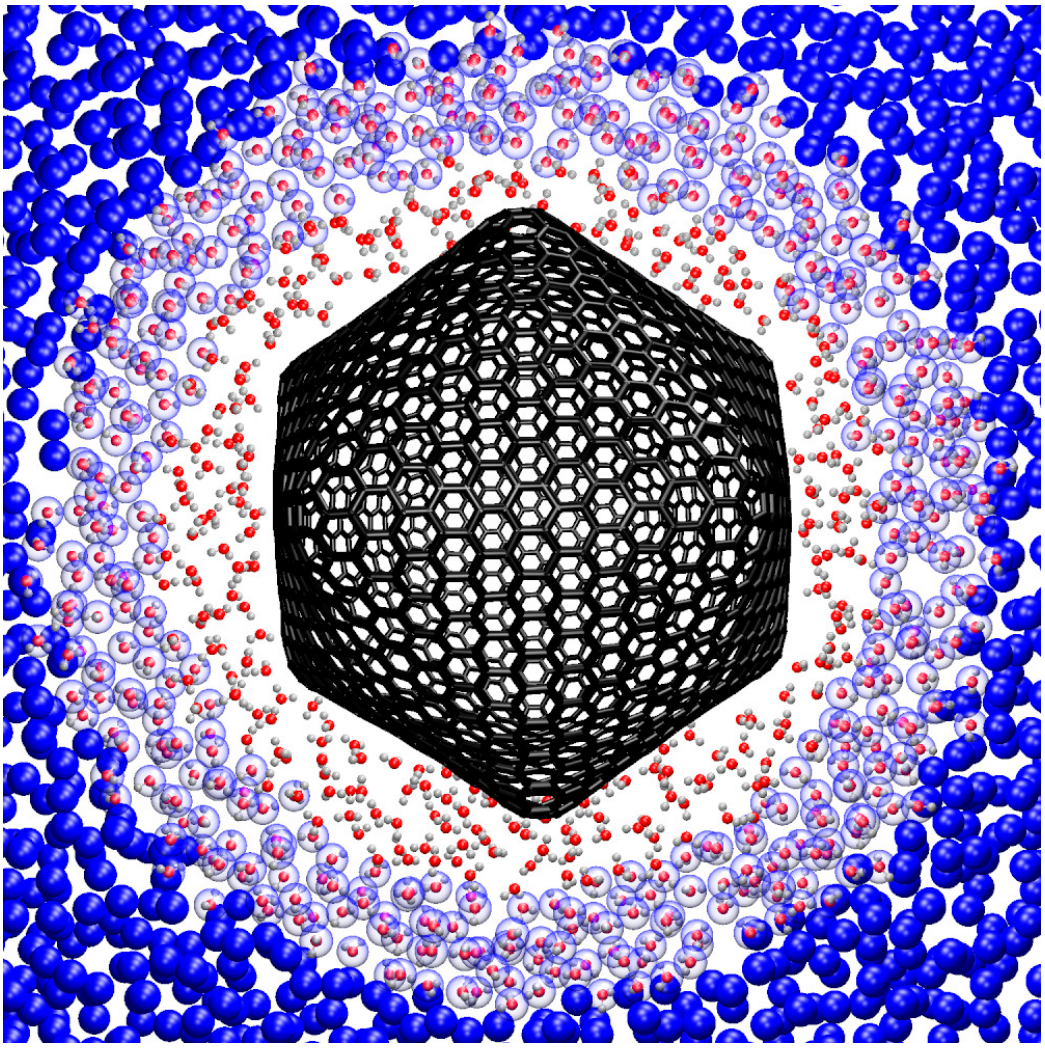


Figure 3.1 : Adaptive resolution simulation scheme for hydrophobic solutes, illustrated for the case of C_{2160} icosahedral fullerene.

To systematically test the locality of hydrophobic hydration, the $60n^2$ series of icosahedral fullerenes (C_{60} to C_{2160}, with corresponding effective radii from 0.35-2.1 nm) were studied by AdResS simulations with varying thickness of the layer of atomistic molecular representation. Fig. 3.1 illustrates the computational setup (see below for details). If adaptive resolution simulations with a minimal all-atom region can

reproduce the results of fully all-atom simulations, then the effect of solute on the water structure is deemed local: it is not significantly supported/influenced by the bulk hydrogen bond network, as the bulk is modeled by CG water, unable to form directional hydrogen bonds. [28, 21] In this case the role of the CG bulk water for the surface layer is minimal, that is, it is enough to ensure that the CG water acts as a thermodynamic bath with the same temperature, pressure and density as the all-atom model. If the structural properties are perfectly local, there is no need for the CG bulk to provide any structural information. Regarding the specific character of the water-solute interaction, the explicit form of the potential has been long debated in the literature. [39, 40, 41] Electronic structure calculations for benzene in water suggest the use of a Lennard-Jones type potential for the C-O non-bonded interaction. [42] Alternatively, the C-O interaction is sometimes treated as a purely repulsive interaction, and modeled with either a Weeks-Chandler-Andersen potential [43], or with a modified LJ potential with the $\left(\frac{\sigma}{r}\right)^6$ term removed. [44]

3.2 Simulation Setup

The application of AdResS to solvation processes requires the definition of an atomistic region around the solute, immersed in a "thermodynamic bath" of CG bulk water. In AdResS two molecules α and β interact with a total force $F_{\alpha\beta}$, evaluated via a space dependent interpolation of the atomistic force $F^{\text{atom}}_{\alpha\beta}$ and CG force $F^{\text{cg}}_{\alpha\beta}$ as $\mathbf{F}_{\alpha\beta} = w(R_\alpha)w(R_\beta)\mathbf{F}^{\text{atom}}_{\alpha\beta} + [1 - w(R_\alpha)w(R_\beta)]\mathbf{F}^{\text{cg}}_{\alpha\beta}$, where $w(R)$ is a switching function;

$w(R)$ is zero in the CG region, one in the spherical atomistic region (see Fig.3.1), and continuous and monotonic in a transition region in between. The all-atom sphere is located at the center of the simulation volume, to which also the center of mass of the solute is constrained, ensuring that the solute is always expressed in full all-atom detail and surrounded by a layer of all-atom water. [27] A 1 nm transition region of hybrid molecules smoothly couples the layer of all-atom water around the solute with a bulk CG water region.

As is evident from Fig. 3.1, the spherical AdResS setup creates an all-atom region of non-uniform thickness around the larger fullerenes. The distribution of the distances of the C atoms from the solute center of mass is spread over an interval Δr, with Δr ranging from less than 0.05 nm for C_{60} to about 0.5 nm for C_{2160}. [45, 46] Nevertheless one can calculate a radial distribution function for the water molecules measured from the surface of the solute and define a first and second solvation shell, even though this is somewhat smeared out for the larger solutes. The distance in nm of the first and second hydration shells measured from the center of mass of the solute are as follows:

Interaction	Shell	C_{60}	C_{240}	C_{540}	C_{960}	C_{1500}	C_{2160}
Lennard-Jones	1st	0.85	1.20	1.53	1.90	2.25	2.55
	2nd	1.05	1.40	1.73	2.10	2.45	2.75
Purely Repulsive	1st	0.95	1.30	1.63	2.00	2.35	2.65
	2nd	1.25	1.60	1.93	2.30	2.65	2.95

where we observe slight deviations between the two surface potentials. The AdResS

setup for the thinnest all-atom layer studied is illustrated in the inset of Fig. 3.2 for the largest solute. The green circle marks the average distance of the C atoms from the center of C_{2160}. The red circle marks the boundary of the all-atom region, which in this case also coincides with the minimum between the first and second peak of the solute-center water $g(r)$. The purple circle marks the border of the transition regime to the fully CG region. For both potentials we study two cases, namely an atomistic region up to the first and second minimum in the solute water $g(r)$, keeping in mind that the position of these minima is slightly different for the two interaction potentials. To actually analyze the solvation shell itself we calculate the number of water molecules $S(d)$, where d is the distance from the solute surface and *not* from the center of the simulation shell (see definition below).

All of the results presented were obtained for NVT simulations using a modified version of the GROMACS package [47, 48]. Stochastic dynamics with a friction constant of $\Gamma = 5$ ps^{-1} and a time step of 2.0 fs was used, and the electrostatics were modeled by a reaction-field method. The volume of the system was obtained from all-atom NPT simulations with P_{ref}=1 atm using the Berendsen barostat. Periodic boundary conditions and minimum image convention were applied in all directions, and a cutoff of 0.9 nm was used for the non-bonded interactions. After warm-up and equilibration of 1 ns, several trajectories of at least 2 ns were collected for each AdResS setup. To ensure that the results are not influenced by the choice of water model, simulations were tested for rigid and flexible versions of the TIP3P and SPC/E point

charge models. [8, 9] The parameters for the all-atom water models are the same as those implemented in GROMACS 4. A single-site isotropic coarse-grained potential was obtained for each water model to match the corresponding O-O $g(r)$ using inverse Monte Carlo iterations with the VOTCA package. [49] The CG interaction potentials are centered on the O atom and are comparable in shape to other single-site water models. [19, 50, 16] This means that the CG model satisfies only a minimal structural requirement of a two body form; the local average tetrahedral structure is not reproduced. [19] Although results are only shown for rigid SPC/E, the general conclusions drawn hold the same for each three-site model, as the detailed simulation results were qualitatively the same for each of the models.

A previously described OPLS model for buckminsterfullerene was adapted for the $60n^2$ fullerenes. [44]. The functional form of the non-bonded potentials was $U = 4\epsilon[(\frac{\sigma}{r})^{12} - a(\frac{\sigma}{r})^6]$ with $a = 1$ for Lennard-Jones and $a = 0$ for purely repulsive interactions. In the case of SPC/E model the force-field parameters were as follows: The other parameters are determined by Lorentz-Berthelot mixing rules. It is important to note that the typical energy of a hydrogen bond (on average 20.41 kJ/mol for SPC/E), is roughly 50 times larger than the optimal C-O Lennard-Jones interaction. [10]

To quantify the change in water structure around a solute, the average number of water molecules $S(d)$ and the average tetrahedral order parameter $q(d)$ are measured as a function of d from the closest C atom of the solute for a bin size of 0.01 nm.

Table 3.1 : Simulation Parameters for Methane and SPC/E water

Atom	σ (nm)	ϵ (kJ/mol)	q (e)	m (u)
C	0.355	0.293	0.0	12.011
O	0.316557	0.650194	-0.8476	15.9994
H	0.0	0.0	0.4238	1.008

Approximations for $g(r)$ (not shown) based on a fit to the volume of each fullerene as a function of their effective radii were computed and match a previous study of hydrophobic clusters. [41] For one single water molecule i, the order parameter q_i is defined as $q_i = 1 - \frac{3}{8}\sum_{j=1}^{3}\sum_{k=j+1}^{4}\left(\cos\psi_{ijk} + \frac{1}{3}\right)^2$, where ψ_{ijk} is the angle formed by the oxygen atoms of two neighbor water molecules j and k with the oxygen atom of molecule i, and the sum runs over the four nearest neighbor molecules of molecule i. [51] The function $q(d)$ is the average q_i over all water molecules at a given distance d from the solute surface. For perfectly tetrahedral systems $q = 1$, while for disordered systems $q = 0$ on average.

3.3 Results

Fig. 3.2 and Fig. 3.3 show the relative ability of each adaptive resolution simulation to reproduce the local density and structure of water around the solute. Comparison of the right and left panels of Fig. 3.2 immediately reveals that while limiting the

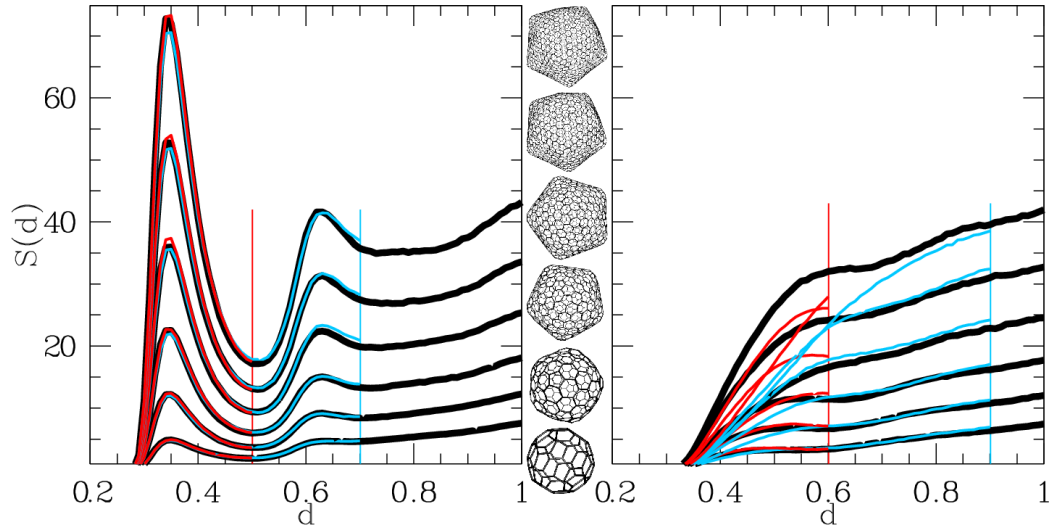


Figure 3.2 : Distribution $S(d)$ of the number of water O atoms at a distance d (in nm, 0.01 nm bin size) from the surface of the solute, for the LJ (left), and the purely repulsive solutes (right). Each set of curves corresponds to the solutes in the order of their cartoon representation. Vertical lines indicate the average size of the all-atom region in each AdResS simulation. Colored curves correspond to different sizes of the all-atom region: up to the first hydration shell (red), or to the second hydration shell (light blue) as measured by $g(r)$. The thick black line represents the result for the reference all-atom simulation.

size of the all-atom region to the first hydration shell is sufficient to reproduce the density of water around the LJ solutes (left panel), the situation is very different for the purely repulsive solutes, where the radial distribution of the water is greatly disturbed by the coarse-graining of the bulk (right panel). We notice an improvement in the results for much larger all-atom regions (not shown), however all-atom region sizes comparable to the first and second hydration shells are insufficient to accurately reproduce the results obtained in fully all-atom simulation. Thus for purely repulsive hydrophobes, structural changes in the bulk directly affect the local water density in the first hydration shell.

This difference between LJ and purely repulsive solutes is also immediately evident from the distribution of the tetrahedral order parameter $q(d)$ around the solutes, shown in Fig. 3.3. The repulsive solutes (right panel) exhibit almost negligible tetrahedral order close to the solute surface. In particular, right next to the repulsive solute surface the packing appears to be increasingly more random with q$\rightarrow$0 for increasing solute size. Surprisingly, this appears rather independent of the thickness of the all-atom layer, indicating the locality of the tetrahedral order, while the water density in the first layers exhibits the characteristics of a property which needs the support of the bulk. On the contrary, the LJ solutes (left panel) present a smaller, albeit significant, decrease in the parameter q (from the bulk value of q$\simeq$0.6 to q$\simeq$0.3-0.4) closer to the solute surface. These results suggest that LJ solutes induce a locally ordered yet very flexible hydrogen bond network, completely consistent with recent vibrational sum frequency spectroscopy results for water. [52] In addition, the height of the first peak of the radial distribution function $g(r)$ (not shown) for the LJ solutes does not appear to decrease for increasing solute size, consistent with the results of a previous study of spherical solutes. [53] On the contrary, for the repulsive solutes a complete "non-wetting" is observed, as the $g(r)$ function for two largest repulsive solutes has essentially no first peak. [54] The difference in non-wetting behavior observed with the two interaction types is fully consistent with previous studies for attractive and repulsive hydrophobic sheets. [36, 55]

These results show how adaptive simulations are capable of shedding some light

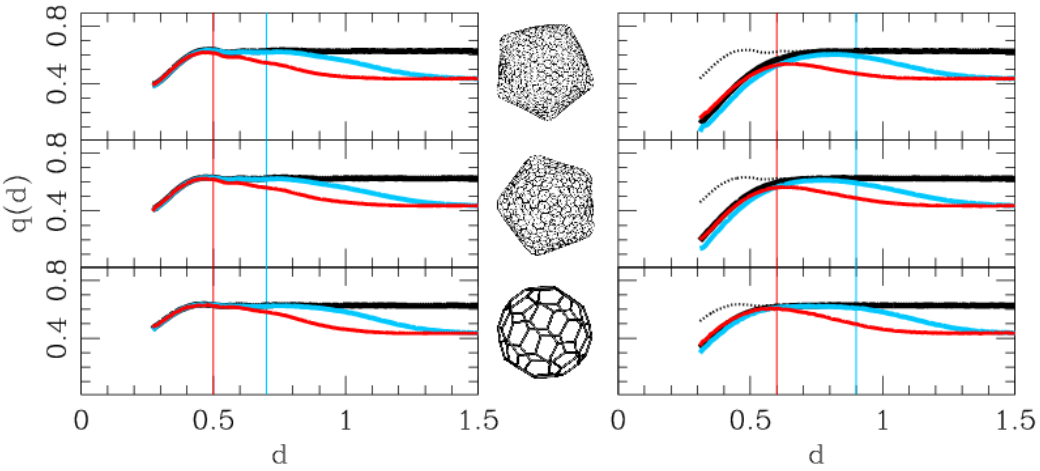


Figure 3.3 : The average tetrahedral order parameter $q(d)$ as a function of the distance d (in nm) from the surface of the solute, for the LJ (left) and purely repulsive solutes (right). Vertical lines and colors are the same as in Fig. 3.2. Results are reported for three representative solutes (see also the corresponding cartoon): C_{60} (bottom plots in both right and left panels), C_{960} (middle plots), and C_{2160} (top plots). On the plots for the repulsive solutes the curves for the fully all-atom simulations of the corresponding LJ solutes are also reported as a dotted black line, to provide a direct comparison of the different cases.

on interface problems. Adaptive simulations that approximate the bulk with a CG model can not match the local density of water around a repulsive solute, while they do for a LJ solute. In contrast, the tetrahedral order in both cases seems to be only weakly affected. As a consequence, contrary to the radial distribution function, the tetrahedral order parameter does not allow for a clear distinction between local and non-local effects for the hydrophobic solute-water interactions as defined in this work.

Although these static quantities are among the most relevant for the structure at the hydrophobic interface, they are not sufficient to determine whether AdResS has adequately captured the essential physics of the solvation process. A complementary dynamic quantity which can support the validity of our conclusions is the particle

Table 3.2 : Density fluctuations near each solute

Interaction	Simulation	C_{60}	C_{240}	C_{540}	C_{960}	C_{1500}	C_{2160}
Lennard-Jones	$AdResS_{d1}$	0.130	0.116	0.116	0.119	0.116	0.115
	$AdResS_{d2}$	0.152	0.140	0.133	0.133	0.130	0.124
	All-atom	0.138	0.131	0.129	0.130	0.146	0.134
Purely Repulsive	$AdResS_{d1}$	0.379	0.643	0.823	1.00	1.22	1.30
	$AdResS_{d2}$	0.585	0.961	1.228	1.45	1.62	1.49
	All-atom	0.463	0.702	0.812	0.911	1.12	1.24

number fluctuation in the first hydration shell. [36, 53, 39] The table below compares the fluctuations, expressed as $\frac{\langle N^2\rangle - \langle N\rangle^2}{\langle N\rangle}$, for each solute in the first layer as defined for Fig. 3.2 and all-atom region size (d1 for red, d2 for light blue). The uncertainty is in the last digit of each reported value. For each LJ solute the adaptive resolution simulations match the all-atom simulation, demonstrating that free exchange of particles occurs between the all-atom and CG regions. A small discrepancy can be observed for the repulsive solutes, reflecting the lack of locality.

3.4 Summary

In summary, the level of locality in the hydrogen bond network at hydrophobic interfaces is primarily controlled by the nature of the interaction between the hydrophobic

solute and water. For purely repulsive solutes, perturbations in the bulk (such as the CG approximation) affect the density but not the tetrahedral order of water at the solute interface. For a weak LJ solute, bulk perturbations tested here do not affect the first water layers. The size of the solute does not appear to affect the locality of the solvation significantly, at least for the range of solutes considered here.

Chapter 4

Chromophore-induced switch in Photoactive Yellow Protein

4.1 Introduction

The postulated signaling state of the bacterial photoreceptor photoactive yellow protein (PYP) is found in experiments to be essentially indistinguishable from an on-pathway folding intermediate formed upon the isomerization of the chromophore embedded in the protein, suggesting a direct link between the folding and the photocycle of this protein. The wide range of time-scales covered during the photocycle/folding of PYP, together with the partially unfolded nature of the signaling state, make the structure determination of the signaling state challenging, either by simulation or by experiment. We have circumvented this problem by using a realistic coarse-grained protein model for PYP, that allows us to characterize the functional changes in the folding free energy landscape. While the folding landscape of the coarse-grained native protein indicates a two-state folding mechanism, the coarse-grained model of the photoactivated form of PYP populates an on-pathway intermediate ensemble, in line with experimental data. The folding intermediate detected in our coarse-grained study matches well all available experimental data on the putative signaling state

of PYP, confirming the connection between the folding and function of this protein. These results allow us to use the coarse-grained model as a starting point to characterize the molecular structure of the signaling state at atomic level of detail. Large structural changes are detected both in the N-terminal domain and in the core of the protein, while its central β-sheet remains largely intact. This study illustrates how the combination of experiment and simulation at different resolutions within a free energy landscape framework can yield novel insight into protein function *.

4.2 Biological Signaling

Biological signaling is initiated when a stimulus converts a receptor into its active signaling state. The accurate characterization of the mechanism of formation of the signaling state is of fundamental importance to understand biological signaling; it is also of medical relevance since erroneous receptor activation, or deactivation, is directly involved in a number of diseases, including impaired vision [56], and cancer [57]. A key step in understanding receptor activation and signaling state formation is the determination of the structure (or ensemble of structures) of the receptor signaling state, and the structural changes that cause relay of a signal from the receptor to its downstream signaling partners. We address these questions for photoactive yellow protein (PYP). Here we show that the long-timescale dynamics leading to the formation of the signaling state can be, at least in first approximation, efficiently

*This chapter is adapted from unpublished work by Ledbetter, Lambeth, and Clementi

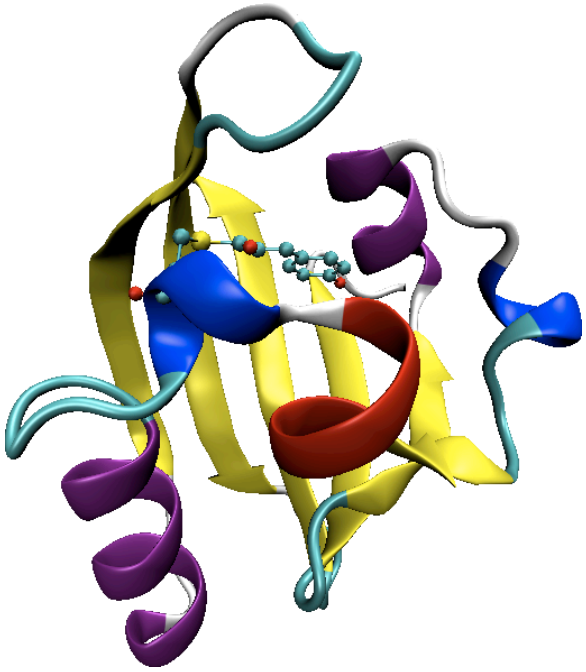


Figure 4.1 : Cartoon representation of the crystal structure of PYP.

simulated by means of a realistic coarse-grained protein model. The coarse-grained study provides a solid starting point for the prediction for the PYP signaling state, as a large ensemble of partially unfolded structures that is found to be in agreement with experimental data.

PYP is a small (125 residues) water soluble protein believed to be responsible for light-mediated negative phototaxis of purple photosynthetic eubacteria [58]. The overall structure of PYP comprises an N-terminal domain containing two α helices (residues 1-25), followed by a typical PAS domain fold (residues 26-125) containing a central six-stranded β-sheet flanked by α helices [59]. Critical to its photoactive capability, the protein contains a deprotonated chromophore, p-coumaric acid (pCA) [60],

which is covalently linked to CYS69. Upon exposure to blue light, pCA isomerizes from *trans* to *cis* form [61, 62], triggering the PYP photocycle.

During its photocycle in solution, PYP undergoes a global structural change [63, 64] that results in the formation of the partially unfolded putative signaling state (pB) [65, 66]. The accurate characterization of the pB state still poses fundamental challenges as the large conformational changes associated to the pB formation occur on the millisecond timescale. A family of NMR structures reflecting the pB state of a truncated variant of PYP (Δ25-PYP) [67] has been reported. However, the complications due to the partially unfolded nature of the pB state has complicated the determination of its structure for the full PYP protein in crystal[66, 65, 67], and NMR spectroscopy has been hindered by loss and overlap of signals, and by the reduced number of structural constraints [68].

Experiments have demonstrated a direct connection between the folding pathway of PYP and its signaling state [69, 66, 65]. In denaturant dilution rapid mixing experiments native PYP containing the *trans* chromophore (pG state) folds in a two-state manner whereas the protein with the chromophore in *cis* configuration populates an on-pathway intermediate during folding [65]. This on-pathway intermediate was found to be kinetically indistinguishable from the putative signaling state structure [65]. Taken together, the experimental findings suggest that the interactions present between pCA and the native protein play a key role in PYP folding. Isomerization of the chromophore disrupts the native interactions during the photo-

cycle [70, 71, 72, 73], thus triggering the global structural changes during the signaling state formation.

4.3 Computational studies of PYP photocycle

In principle, molecular simulation methods such as molecular dynamics (MD) could elucidate the structural changes that take place during the PYP photocycle/folding. Many preceeding molecular dynamics and QM/MM studies have developed significant insight into the structure changes in PYP up to the nanosecond timescale after photoactivation [74, 75, 76, 77, 78, 79]. However, simulation beyond the millisecond timescale, where convergence to the signaling state is known to occur, is problematic due to the number of degrees of freedom, and the large number of intermediate traps inherent in protein landscapes.

One direction for bypassing the long timescale limitation in PYP, replica exchange simulations [80] and transition path sampling [81] have recently been used to gain insight on the pB structure. A complementary alternative, presented here, is to use a coarse-grained protein model.

Coarse grained models like the familiar Gō-model, and the DMC model used here, combat the timescale difficulty on both fronts. In both models, the number of degrees of freedom in the model are reduced by an order of magnitude by grouping the atoms of each amino acid into a single coarse grained atom. The long range interactions between the amino acids are approximated by a computationally cheaper energetic

potential as a long range interaction between the coarse grained atoms. Then, in the Gō-model, by the assumption of the minimum frustration principle, we assume that any residue contacts not found in the native geometry have neglible stability, and hence attractive interactions between non-native pairs can be neglected without compromising the integrity of the long timescale dynamics.

Gō-like coarse-grained models have been widely used in studies of protein folding [82, 83, 84]. However, a drawback of the Gō-model is that, by design, only a single, non-flexible stable state can be represented. Unfortunately, when a protein's function requires reliable and repeatable transformation among one or more distinct geometries, one must go beyond the unfrustrated/single-basin Gō-model. This requirement is evident in the nature of light transducing signal proteins: evolutionary fit operation of such proteins logically require two stable geometries. First, the signal cascade *should not* be catalyzed in the absence of the photo stimulus, and second, that the protein reliably catalyzes the signal cascade after photo-activation. Such a mechanism necessitates two distinct geometries.

The limitation of a single stable geometry in the Gō-model can be overcome if *a priori* information about the alternative/functional geometries is known. For example, if the alternative geometry is known, the two geometries can be incorporated into a multi-basin Gō-model. In the case of PYP this option is not available since the geometry of the signaling state is not experimentally known. Instead, we use the experimentally known behavior, in conjunction with energetic frustration, to predict

the alternative geometry produced by photo-activation. By itself, knowledge of the switching behavior in the chromophore is not sufficient to develop a prediction of the signaling state in the Gō-model since the non-native contacts in the active form of the protein can not be stabilized. We demonstrate this deficiency of the Gō-model by studying the free energy landscape of the coarse grained model after photo-activation.

In order to predict the signaling state of PYP, we use the coarse grained DMC model to combine the knowledge of the chromophore induced switch in PYP, with the information about energetic frustration implicitly contained in the amino acid sequence. This is accomplished by developing a DMC model for chemically inactive dark state, and photo-activating the model by perturbing the amino acid interactions involving the chromophore.

While a two-state folding mechanism is observed for the coarse-grained model of PYP, consistent with experiments, the folding mechanism of the photoactivated form is found to be three state, in agreement with the experimental data. On the contrary, results obtained with a Gō-like coarse-grained model do not present any evidence of any alternative stabilized geometries when the model is perturbed by photoactivation (see Figure 4.2).

The folding intermediate ensemble detected on the folding landscape of the DMC model of photoactivated PYP reproduces all comparable experimental observations. In particular, we consider comparisons of the kinetic and structural characteristics of the predicted pB state to experimental results. The agreement enables us to predict

the main features of the signaling state of PYP, and demonstrates that the disruption of the native pCA-protein interactions greatly alters the folding free energy landscape of PYP. The ensemble of coarse-grained structures obtained with the DMC model provides us the starting point for a more detailed all-atom characterization of the PYP photoactived state.

4.4 Model and Methods

4.4.1 Coarse-grained protein model

PYP is modeled using the DMC off-lattice minimalist representation in which each amino acid is described by a single bead on a polymer chain located at its C_α position. The potential energy of a protein configuration comprises a local and a non-local term, $V = V_{local} + V_{non-local}$, where V_{local} includes bond, angle and torsional energy terms and is designed to have its absolute minimum in the native state, as in previous work [83, 84]. However, the details of V_{local} are not crucial for the results obtained in this study; a statistical potential could be used in its place.

The non-local potential energy function for a protein of N residues describes the non-covalent interactions between all residues separated by at least three residues along the chain, and is defined as:

$$V_{non-local} = \sum_{\substack{i,j=1 \\ i<j-3}}^{N} \epsilon(c_i, c_j) \left[5 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \delta(c_i, c_j)\, 6 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{10} \right]. \tag{4.1}$$

Each residue is associated with a specific amino-acid type or "color" c_i, that

differentiates its physical and chemical properties. The parameter $\epsilon(c_i, c_j)$ determines the strength of an effective non-bonded interaction between a residue pair (i, j), and is set to be positive. The parameter $\delta(c_i, c_j)$ determines if the non-bonded interaction between a residue pair (i, j) is attractive ($\delta(c_i, c_j) = 1$) or repulsive ($\delta(c_i, c_j) = 0$).

The equilibrium distances σ_{ij} for non-local interactions carry information on the geometric complementarity of the interacting residues (which would be lost if a constant value of σ_{ij} were used). The parameters σ_{ij} for native PYP are defined in the following way: For each residue pair (i, j), if at least one heavy atom of residue i is closer than 4.5 Å to a heavy atom of residue j in the protein crystal structure (pdb code 2PHY for native PYP), the interaction is considered to be native and σ_{ij} is set equal to the distance between the C_α atoms of residues i and j in the native state. For all other non-local interactions, σ_{ij} is randomly extracted from a distribution $P(\sigma; c_i, c_j, |i - j|)$, as explained in detail in ref [85]. We consider three different distributions for each pair of residues (c_i, c_j), according to the relative distance between residues i and j along the sequence: The cases $|i-j| = 4$, $|i-j| = 5$ and $|i-j| > 5$ are considered separately. The distribution $P(\sigma; c_i, c_j, |i - j|)$ is obtained by performing a statistical analysis on the $C_\alpha - C_\alpha$ distances over all native contacts formed by pairs of residues of types c_i and c_j, with the desired distance $|i - j|$ along the protein sequence, from a large (> 4000) database of non-redundant protein structures.

4.4.2 Definition of an optimal set of parameters to model native PYP

The effective energy parameters, ϵ and δ (see Equation 4.1), for a given amino-acid sequence are designed in such a way that the protein is able to fold to its native state. The parameters $\epsilon(c_i, c_j)$ and $\delta(c_i, c_j)$ for native PYP are generated through an iterative procedure, described in detail in ref [85]. The optimal set of parameters $\{\epsilon_{opt}, \delta_{opt}\}$ is defined by the maximum energy gap criterion [85, 86, 87] on a large set of decoys (i.e. low energy, compact, non-native structures) such that $\Delta E(\{\epsilon_{opt}\}, \{\delta_{opt}\}) = \max_{\{\epsilon\},\{\delta\}} \left[\left(\min_{i\in(1,\ldots,N_{decoys})} E_{\Gamma_i}\right) - E_{\Gamma_{nat}}\right]$, where $E_{\Gamma_{nat}}$ refers to the energy of the crystal structure Γ_{nat} of native PYP and E_{Γ_i} is the energy corresponding to the decoy structure Γ_i. Any low energy configuration with RMSD > 2.5 Å from the crystal structure is considered a decoy structure. An initial set of 1000 decoys is used and a total of 110 iterations is performed. Multiple short "heat-and-quench" unfolding/refolding simulations are carried out to check if the obtained set of parameters is able to refold the protein to the correct native state (within RMSD < 2.5 Å from the crystal structure of native PYP). A total of ~ 4000 decoys are used to obtain one set of parameters that is able to consistently fold native PYP back to its native state starting from an unfolded configuration generated at high temperature. This foldable set of parameters is modified to model photoactivated PYP, as discussed in the next section.

4.4.3 Coarse-grained model of photoactivated PYP with disrupted chromophore-protein interactions

The DMC coarse-grained model used in this study cannot explicitly describe the important structural details of the chromophore and its binding pocket (for example, the strain in the chromophore and electrostatic effects related to the proton transfer between pCA and the active site residue GLU46 [88, 89]) to model the folding mechanism of PYP with *cis* chromophore. However, recent experimental [71, 90, 91, 72, 73] and theoretical [92, 77] studies indicate that the main effect of both the pCA isomerization and proton transfer is to disrupt the chromophore-protein interactions present in the native PYP. This observation allows us to design a "photoactivated" form of PYP (which can be thought of as a coarse-grained equivalent of PYP containing *cis* chromophore) by effectively removing the native interactions between pCA and the protein in the coarse-grained model. The disruption of native interactions introduces destabilization to the native state of PYP. This effect is obtained by performing the following modifications on the parameters associated with non-bonded interactions between CYS69 and any other residue: (i) We set $\delta(c_{69}, c_j) = 0$ for every residue $j \in (1, \cdots, N)$ in order to maintain only the repulsive core for all the interactions with CYS69 in photoactivated PYP; (ii) We keep the value of the parameter $\epsilon(c_{69}, c_j)$ for photoactivated PYP the same as for native PYP, for every residue $j \in (1, \cdots, N)$. All the interactions that are present between the chromophore and any other residue in native PYP are considered non-native for photoactivated PYP and the equilibrium

distances σ_{ij} for such interactions are randomly extracted from the corresponding distribution $P(\sigma; c_i, c_j, |i-j|)$. The main approximation in this model of photoactivated PYP is that all interactions of residue 69 are considered purely repulsive. The fact that the free energy of the pB_{calc} state is found ~ 1 RT_f higher than the pG_{calc} state can be interpreted as a consequence of this approximation.

4.4.4 Folding thermodynamics and kinetics

The set of parameters defined in the previous sections are used to characterize the folding of native and photoactivated PYP. We perform several long folding/unfolding MD simulations with around the folding temperature T_f for native and photoactivated protein using the GROMACS 4.5 package [47]. Simulations at different temperatures are combined using the WHAM algorithm [93, 94] to obtain the folding free energy surfaces projected onto a number of different reaction coordinates. The folding temperature T_f is estimated as the temperature corresponding to the peak of the specific heat profile as a function of temperature.

4.4.5 Determination of reaction coordinates

The LSDMap procedure [95, 96] uses Principal Component Analysis on varying sizes of local neighbors to determine the local scale of each cluster. These local scales are then used to determine a Diffusion Map for the trajectory, which produces approximate eigenvalues and eigenvectors for the Fokker-Planck diffusion equation. The coordinates of each configuration in this eigenvector remapping allow for calculating

thermodynamic properties, specifically free energy, in terms of ensemble probability. The relative eigenvalues correlate with the time scale of the motion captured by the respective eigenvector.

A free energy associated with different stages of the folding reaction are obtained by taking the final configuration of multiple simulations of folding trajectories longer than the folding time for the models. Figure 4.2 shows the folding free energy surfaces as obtained from simulations for both the native and photoactivated protein as a function of reaction coordinates obtained with the locally scaled diffusion map (LSDMap) approach [96, 95]. The LSDMap method provides reaction coordinates which are ideal for separating diffusion in the free energy landscape according to the timescale of the motion (for instance, separating motion towards the signaling state from the longer timescale folding behavior). For comparison, the results obtained for both the DMC and a Gō-model are presented in Figure 4.2.

The free energy surfaces (Figure 4.2(a) and (c)) indicate that both in the DMC and Go-like models a single preponderant free energy barrier separates the folded and unfolded states of native PYP, in agreement with the two-state folding behavior observed experimentally for this protein [65].

Photoactivation of the DMC model of PYP dramatically affects its folding landscape, creating an on-pathway folding intermediate ensemble (as shown in Figure 4.2(b)). The prediction of this ensemble is in agreement with experimental data; Hoff et. al., have observed an intermediate state during the folding of PYP containing *cis*-

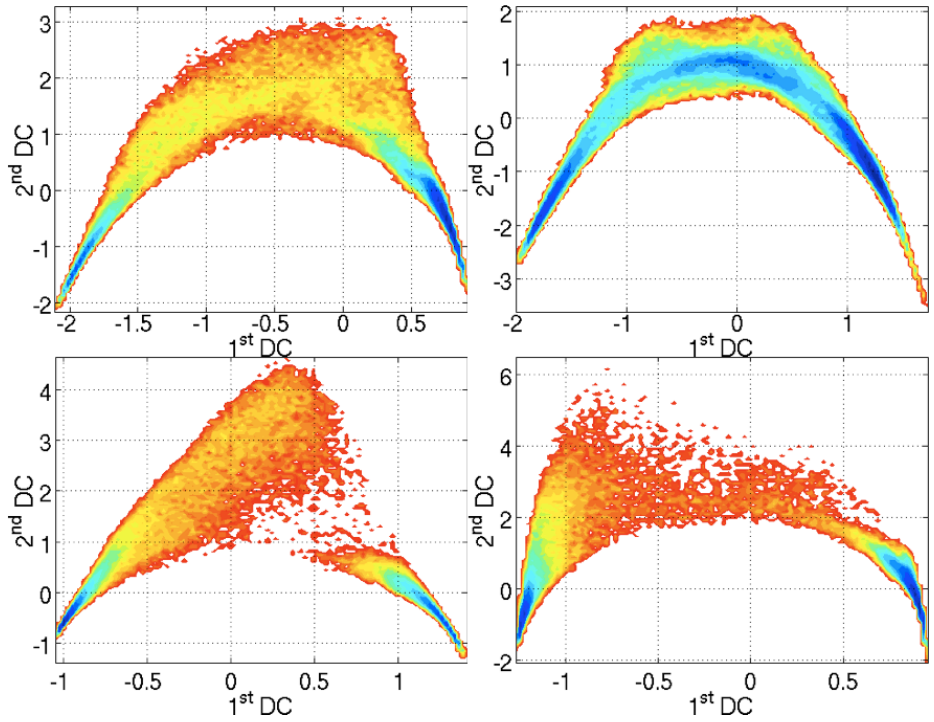


Figure 4.2 : Free energy surfaces for the native PYP (left) and the photoactivated PYP (right) at the corresponding folding temperatures as a function of the first and second diffusion coordinates Top: Results from DMC coarse-grain model. Bottom: Results from Go-like coarse-grain model. Each contour on the free energy surface represents 1 RT_f free energy difference.

pCA [65]. This observation suggests that the photoactivation of the PYP protein in coarse-grained DMC model mimics the changes on the folding behavior detected in experiments as a result of chromophore isomerization.

Although a coarse-grained protein model based on Gō-like potential can correctly reproduce the two-state folding behavior of native PYP, the photoactivation-induced switch in the folding landscape of PYP can not be captured using a Gō-like potential (see Figure 4.2(d)). This observation underscores the role of sequence details and energetic frustration in shaping the free energy landscape governing the folding and

photocycle of PYP.

Denaturant dilution rapid mixing experiments indicate that the on-pathway folding intermediate of PYP containing *cis*-pCA is indistinguishable from the signaling state of PYP [65]. As the modeled photoactivation of the DMC model of PYP describes the perturbation in folding landscape upon pCA isomerization in agreement with experimental data, the simulated folding intermediate ensemble of photoactivated PYP serves as an excellent candidate for representing the experimentally detected pB signaling state. A detailed analysis of these results is presented in the next section, where the pB ensemble detected in the coarse-grained simulation is compared with the results from experimental studies. Supported by the agreement with the available experimental data, the pB ensemble obtained with the DMC model can then be used as a starting point for a finer characterization of the pB state, to the level that can be experimentally tested.

We refer to the coarse-grained folding intermediate ensemble as pB_{calc}. The corresponding folded ensemble obtained for the native protein is referred as pG_{calc} throughout the rest of this chapter.

4.5 Validation of pB_{calc} as the signaling state of PYP and new insights on the signaling state structure

A number of spectroscopic experiments provides valuable information on various aspects of the PYP signaling state [65, 63, 64, 97, 98, 99, 100, 101, 102]. However, a

complete and accurate determination of the molecular structure of pB still remains an open challenge for existing experimental techniques. The pB_{calc} ensemble obtained in our simulation provides us with an excellent candidate for the structural characterization of pB, and we have used the available experimental data on pB state to validate the pB_{calc} structure as a model for the signaling state of PYP. As described below, the quantitative comparison of the structure of pB_{calc} and the pB state detected in spectroscopic experiments shows that the folding intermediate state detected in our simulations is in good agreement with all available experimental results.

Spectroscopic measurements indicate that the putative signaling state of PYP has reduced tertiary structure but largely intact secondary structure [66]. In order to determine the amount of secondary and tertiary structure in pB_{calc} we have examined the probabilities of short-range and long-range native contact formation in the corresponding ensemble of structures. We define short-range and long-range interactions as follows: any native contact between residues i and j is considered a short-range interaction if $|i-j| \leq 20$, otherwise, it is considered long-range. We find only $\sim 40\%$ of the long-range native interactions present in pB_{calc} whereas $\sim 70\%$ of short-range native interactions remain formed. This observation indicates a significant disruption of tertiary structure in pB_{calc} while most of the secondary structure is intact, in agreement with experiment [66]. The reduction to $\sim 40\%$ of the total number of non-local interactions observed pB_{calc} compares well to the decrease to $\sim 45\%$ in the number of mid- and long- range interactions for the pB state that has been mea-

sured in NMR spectroscopy for Δ25-PYP, a truncated form of PYP lacking the 25 N-terminal residues [67].

The regions with lower probability of native contact formation in pB_{calc} roughly fall into two clusters, namely in the N-terminal region and the region around the chromophore. These two regions of the protein correspond to the two experimentally observed hydrophobic patches that become more accessible to fluorescent dyes upon formation of the putative signaling state [103, 66, 98].

The available experimental data enable us to perform a residue-level analysis on the structural displacement and flexibility of PYP upon formation of the signaling state. A number of experiments such as H/D exchange rate measurements and ^{15}N HSQC spectrum have provided insight on the structure of the pB state of the wild type PYP [98, 99, 104]. In addition, recent H/D exchange and NMR experiments on Δ25-PYP have presented a more detailed structural and dynamical description of the signaling state of this PYP variant [67].

In summary, all of these experiments suggest that a number of residues become structurally disordered upon formation of the pB state. In addition to the N-terminal domain, regions involving residues 42-58, 63-78 and 96-103 show enhanced flexibility in the signaling state of PYP. Because of the partially unfolded dynamic nature of the pB state, the NMR spectral features do not allow a detailed structure determination of pB, as the NMR data represent a dynamic average over a large conformational ensemble [105]. However, the sparse NOE information complemented with the change

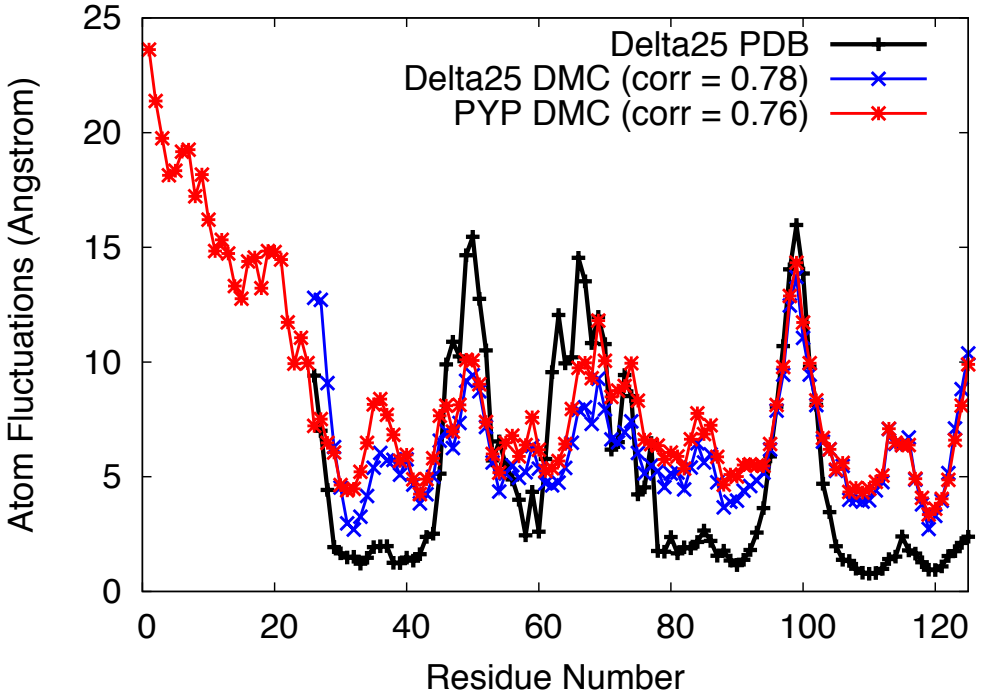


Figure 4.3 : Average displacement (in Å) per residue in the Δ25 pB_{calc} state (blue dots) and in the full PYP pB_{calc} state (red dots) as obtained from our study. Black dots indicate the per residue deviation from the native structure averaged over the 20 NMR-based pB structures of PYP lacking the first 25 residues. The high correlation (r≈0.8) between the average deviation per residue observed in the NMR pB structures and in pB_{calc} ensemble (both of the full PYP and Δ25-PYP) indicates that the DMC coarse-grained model is able to correctly reproduce the main conformational changes in PYP upon signaling state formation.

in HSQC spectra provides a model for the signaling state structure of the pB state of Δ25-PYP. In the NMR-based pB ensemble, secondary structure formation is reported for $\beta 1$, $\beta 2$, $\alpha 3$, $\alpha 5$, $\beta 4$, $\beta 5$, and $\beta 6$ [67, 105]. Most of these secondary structures are found to be shorter in the signaling state in comparison to the ground state of PYP [67, 105].

In order to quantify the structural disorder and the conformational flexibility in

pB_{calc}, and to quantitatively compare it with the experimentally proposed pB state, we have computed the average displacement per residue in the pB_{calc} ensemble. The comparison is presented in Figure 4.3. To make sure that the results obtained on the full PYP protein are comparable with the results for Δ25-PYP, we have repeated the simulation and analysis for a DMC model of the Δ25-PYP, obtained by deleting the first 25 residues in the DMC model of the full PYP (and mantaining the same energy function for the remaining residues). Results obtained for the DMC model of both PYP and Δ25-PYP are compared to the NMR results for Δ25-PYP in Figure 4.3. The red dots in Figure 4.3 mark the deviation per residue in the pB_{calc} ensemble from the native structure, as obtained for the full PYP protein, while the blue dots correspond to the pB_{calc} ensemble of Δ25-PYP.

In both the full and Δ25 version of PYP, a significant part of the protein undergoes structural modification in the pB_{calc} ensemble, with regions of profound structural alteration corresponding to residues 1-26, 45-52, 56-60, 64-78, 94-106 and 124-125. The flexibility of the protein in the pB_{calc} ensemble can be evaluated by computing the rmsd for residue from a representative pB_{calc} structure, over all the conformations belonging to pB_{calc}. Figure 4.3 shows that the flexibility plot (purple dots) closely reproduces the behavior of the displacement plot (blue dots), clearly indicating that structural differences between the pB_{calc} ensemble and the native structure is mainly associated to the partial unfolding of limited regions of pB_{calc}. Interestingly, a similar result was obtained in a previous study on the effect of mutations on S6 protein: The

interplay between partial unfolding and increased flexibility was found the main factor in determining key structural changes in the S6 native ensemble upon a particular set of mutations. While for PYP the partial unfolding of certain regions may be exploited for functional reasons, in the case of S6 these structural changes are likely responsible for the increased aggregation propensity of the mutant [106, 107].

The regions of pB_{calc} (both for the full and Δ25 PYP) that show higher structural displacement (and enhanced flexibility) match very well with the regions that are found in experiments to be poorly structured upon formation of the PYP signaling state [98, 99, 104, 67, 105]. In particular, the per residue deviation observed in pB_{calc} compares well with the per residue deviation measured over the 20 available NMR structures of the pB state of Δ25-PYP (black dots in Figure 4.3). The average displacement per residue in the NMR-based pB ensemble is $\sim$ 4.8 Å compared to $\sim$ 5.3 Å in pB_{calc}.

The calculated per residue displacement in pB_{calc} is strongly correlated (Pearson Correlation coefficient $r \sim 0.8$, both for full and Δ25 PYP) to the corresponding displacement obtained for the NMR ensemble of Δ25-PYP (black dots), and the displacement curves present the same features. The agreement is particularly remarkable given the approximations used in the model, and the fact that the NMR ensemble of Δ25-PYP is composed by only 20 structures, as opposed to $\sim$ 1000 structures of pB_{calc}. The helical regions in pB_{calc} experience a large structural perturbation and enhanced flexibility compared to the β-strand regions, in agreement with the exper-

imental findings [99, 98, 67] and consistent with the results obtained from previous computational studies [92].

As discussed in the *Model and Methods* section, our model underestimates the contribution of interactions involving the *cis*-chromophore in the pB state and, as a result, pB_{calc} is found $\sim$ 1 RT_f higher in free energy than pG_{calc}. These interactions may strengthen the non-native contacts involving the chromopore and the chromophore-containing loop in the pB state and tip the free energy balance in favor of the pB state.

Taken together the results presented in this study show that a realistic coarse-grained protein model can successfully provide a solid starting point for the characterization of the detailed molecular structure of the signaling state of PYP, a result that remains challenging for NMR experiments and standard MD all-atom simulations alone. For this reason, we refine the results presented here at atomic level of detail and we find that, indeed, a more accurate energy function further stabilizes the folding intermediate ensemble with respect to the native state.

4.6 From coarse grain to all-atom

Because of the large range of timescales involved, standard MD simulations are restricted to the characterization of the initial events in the photocycle. Previous computational studies of PYP have used mixed quantum mechanics/molecular mechanics (QM/MM) simulations [108, 76, 109] to characterize the fast processes of the PYP

photocycle. In order to accelerate the sampling of metastable states, replica exchange and transition path sampling have been used to obtain a qualitative description of the pB state [80], the mechanism leading from pB' to pB state [81], and the recovery of the pG state [110].

In parallel to the development of advanced sampling methods, multiscale techniques are emerging as promising tools to combine the efficiency of coarser simulations with the detail of all-atom simulations for the characterization of a broad range of macromolecular systems over long timescales. During last decade several methods have been proposed to combine different resolutions in different regions of the conformational space during a single simulation [111, 112, 113], or to change the resolution over the whole system during the course of a simulation [114, 115, 116]. Alternatively, multi-stages approaches have been proposed, in which simulations at a coarser resolution enable calculations at finer resolution. One of the first examples of multi-stage simulations for protein systems used a reweighting scheme to estimate all-atom free energy from the conformational sampling obtained by means of a simplified representation of a 12-residue helix [117]. In a more recent study, marginally compact structures generated by means of Monte Carlo simulations of a coarse grained model were taken as viable initial configurations for extensive all-atom MD of the folding of the villin headpiece protein [118].

The idea of using a coarser resolution for a fast approximate exploration of a complex configurational landscape, that can then be used as a starting point for

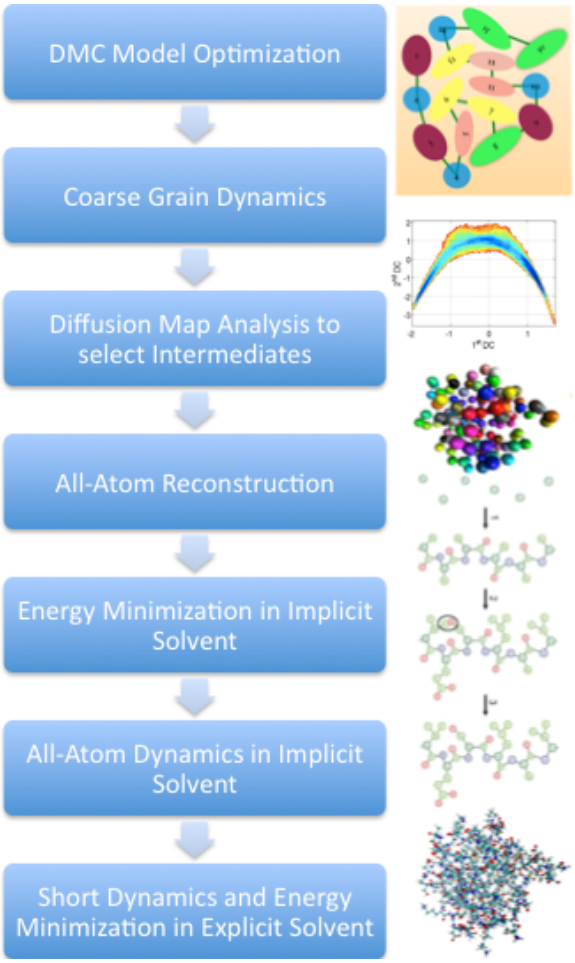


Figure 4.4 : Flowchart for Prediction of pB_{calc} signaling state

studies at a finer resolution, has also been applied in different contexts. For instance, a recent example is the exploration of the conformational space of peptides with classical MD, that has been used as a starting point for structure prediction with density functional theory [119, 120]. Another example is the use of coarse-grained MD simulations backmapped to all-atom resolution to investigate the structure of polymeric melts [121, 122].

Recently a fast and reliable reconstruction procedure has been proposed that allows to systematically reintroduce the atomic degrees of freedom associated with a C_α-only coarse-grained protein configuration [123]. In brief, the backbone atoms are first introduced by using the results from a statistical analysis (on a large non-redundant set of protein structures) of the relative position of the backbone atoms associated with different configurations of four consecutive C_α atoms. The side chain atoms are then reintroduced by discretizing the possible side chain orientations of each residue in a set of backbone-dependent rotamer states [124], and selecting the rotamers associated with low energy configuration. (see [123] for detail). Here we use this all-atom reconstruction procedure as part of a more general scheme to obtain a prediction of the pB state of PYP. The scheme is illustrated in Figure 4.4. Instead of reconstructing to all-atom resolution the whole conformational landscape visited during the coarse-grained simulations (as was done in previous work [123]), here we focus our attention to the accurate comparison between the pG and pB states. We assume that the crystal structure of PYP in its native form provides a good representation of the pG state. For a refined characterization of the pB state, we start by selecting the 25 "most probable" configurations from the corresponding coarse-grained ensembles, and reconstruct each of them to all-atom resolution.

In order to explore the conformational space around the reconstructed structures, we use these structures as starting points for all-atom MD simulations. From each structure, 100 independent trajectories are started and evolved for 30ns by using the

AMBER99 force-field in implicit water. This diffusion dynamics, albeit short, provide information on the landscape around the selected structures [125]. The final structures of each trajectory in the diffusive dynamics are then solvated in explicit water and minimized in the presence of counterions. These solvated, minimized structures constitute our prediction of the pB state of PYP. Figure 4.5 shows the lowest energy structure from this ensemble.

The solvation procedure is particularly important because of the stabilizing effect it provides for the intermediate structures. Water mediated interactions are not captured in the coarse grained model, and relative energies based on implicit solvation also fail to capture the multi-body effects of solvation. Recent studies of src-homology 3 (SH3), the canonical two-state folding protein, have shown multiple parallel pathways for folding [126, 127, 128]. This observation has been shown to be difficult to capture with a vacuum coarse grained model [129, 85] but is unavoidable when coarse grained solvent is included in the system. Therefore, it is important that explicit solvent is used for comparing the energies of the pB and pG structures of PYP.

The main underlying assumption of this approach is that the coarse-grained model used for the initial exploration must be realistic enough that the protein structures being sampled implicitly represent relevant conformations of the protein. This assumption can not be rigorously proved *a priori*, and we can not completely rule out the possibility that the ensemble of the most probable structures obtained in the coarse grained model completely misses the regions that would be relevant for an

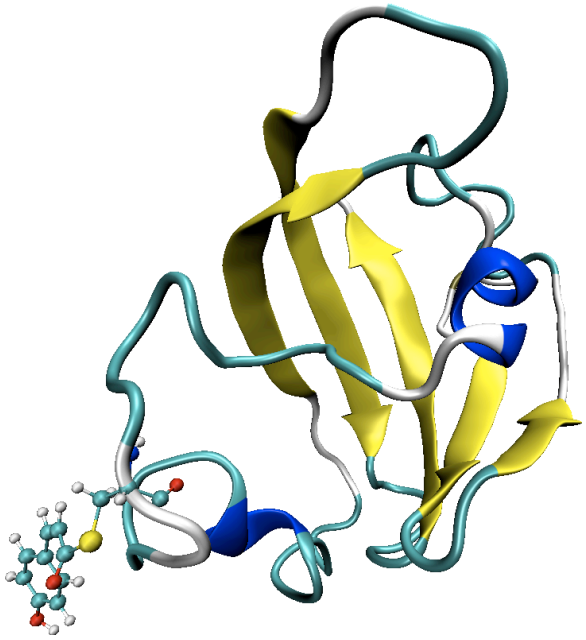


Figure 4.5 : Cartoon representation of the predicted pB photoactivated state. The chromophore pocket is open and flexible while the beta sheets are maintained. The structure is in close agreement to other predictions from replica exchange MD and the Δ25 NMR structure.

all-atom model; however, as discussed in the companion paper, the good agreement of the average properties of the coarse-grained pB ensemble with all available experimental data supports the idea that this pB ensemble has a significant overlap with the region of the conformational space corresponding to the "true" pB state. Ultimately, only the experimental testing of the predicted pB state can assess the validity of this approach.

4.7 Predicted signaling state ensemble at all-atom resolution

To our knowledge, there is only one previous computational prediction of the signaling state of PYP, that was obtained with replica exchange MD. The representative structure of our pB ensemble shown in Figure 4.5 is very similar to the representative structure of PYP presented in Fig.8 of [80]. In particular, the chromophore is completely exposed to the solvent, and the N-terminus of the protein appears mostly unstructured, as well as the chromophore pocket, while the β-sheet remains well structured. Interestingly, both in our predicted pB state and in [80] the β-sheet appears more "flat" than in the pG state, where it is curled, as to cover the chromophore pocket. The persistence of the β-sheet in the pB state, with the rest of the protein fluctuating (to different degrees in different regions) around it is consistent with the experimental observation that PYP in its pB state can readily form amyloid structures, while the formation of such structures is not observed in the pG state.

4.8 Conclusions

In summary, the coarse-grained protein simulation study successfully detects the signaling state on the folding landscape of PYP, a partially unstructured structural ensemble, that is not accessible to either all-atom simulations or existing experimental techniques alone. The results presented in this work provide strong evidence in support of the hypothesis (formulated on the base of experimental evidence) that the free energy landscapes for folding and function are directly linked in PYP. The

interplay between folding/flexibility and function is likely to apply to a broader range of proteins, where structurally disordered local minima populated during the folding can act as functionally active states for a protein [130, 131, 132, 133, 134, 135]. The results presented here underline the importance of realistic coarse-grained protein models in addition to experiments and fully detailed simulation studies to achieve a deeper understanding of the the factors determining biological functions on a protein landscape. Such coarse-grained protein models can provide a solid starting point to build a multi-scale theoretical/simulation framework [136, 137, 138, 139, 140] that can be complemented with refined simulations or experimental techniques for a more quantitative understanding of a protein folding-function relationship. The application of such multiscale techniques could be used as a practical diagnostic tool to study biophysical processes such as misfolding or "protein quakes" where a multi-scale description of a protein system is needed to bridge the gap in time-scale and length-scales between atomic scale simulations and wet lab experiments.

Chapter 5

Coarse Grained Protein in Coarse Grained Solvent

5.1 Introduction

The bottleneck in studying protein dynamics lies in simulating the slower degrees of freedom while simultaneously minimizing the loss of the cumulative effect of minute, faster degrees of freedom. This is a pervasive problem across multiple computational areas in material science and biophysics, for which several different strategies have been proposed in the last few decades. Methods like transition path sampling [80] and temperature accelerated molecular dynamics [141, 142] allow to improve the sampling of metastable states. Another popular approach to this issue has been to develop coarse grain (CG) models [143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 17, 158, 19] to reduce the necessary computational time for adequate sampling. Several methods have been proposed to combine different resolutions in different regions of the conformational space during a single simulation [111, 112, 113], or to change the resolution over the whole system during the course of a simulation [114, 115, 116]. Alternatively, multi-stage approaches have been proposed, in which simulations at a coarser resolution enable calculations at finer resolution [117, 118, 119, 120, 121, 122].

One area of recent research has been the extension of coarse graining to the solvent. The majority of existing coarse grain protein models do not explicitly resolve the solvent, which does not allow for investigation of water mediated interactions. To the best of our knowledge, no existing single-site water model has been used to study long time-scale protein dynamics involving large structural rearrangements such as a folding process. There are only a few instances of CG water models for biomolecular simulations [159, 113, 160, 150, 155, 17, 161, 162, 19], none of them directly applicable to the study of water mediated processes due to the resolution of the models: they all consider 3 or 4 water molecules as a single-site, which is not capable of penetrating the protein core due to it's collective size.

By using a previously developed CG water model [28, 21], and starting from the DMC protein model [85], we have succeeded in the definition of a set of water-protein and inter-protein parameters able to consistently and smoothly fold/unfold/refold a CG model of SH3 in a "solution" of CG water. The results suggest the presence of a "water-expulsion" process after the collapse transition state, where a few water molecules need to be expelled from the protein core in order to proceed to the completely folded structure. This is in agreement with previous studies at different resolutions [129, 163], encouraging the use of explicit CG water in the study of long time-scale protein dynamics.

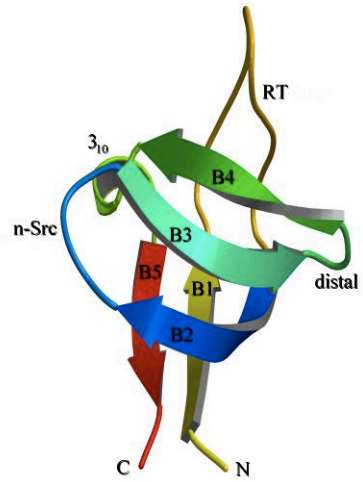


Figure 5.1 : Cartoon representation of the crystal structure for src-SH3. The structure consists of two orthogonal beta sheets joined by an RT loop and a diverging turn.

5.2 Models and Methods

5.2.1 Coarse Graining

SH3 is modeled using the DMC off-lattice minimalist representation [85, 164], described in Chapter 4, which is based on the minimization of frustration in a C_α-only model subject to the constraint that all residues of a particular amino-acid type exhibit the same non-local interactions. This requirement naturally introduces energetic frustration, which is minimized by requiring the energy gap between the native structure and alternative compact misfolded structures to be as large as possible. The parameters of the model are optimized to fold into a known protein structure; in this sense, the approach maintains the spirit of structure-based models [83, 129, 165, 166]. However, by modeling the protein energy as a sum of pairwise interactions between

20 amino-acid "colors", the DMC model does not *a priori* remove the contribution of non-native interactions, and naturally incorporates the energetic modulation of a protein sequence. The simplified Hamiltonian effectively reproduces the mean effects of non-local interactions between the residues over long time-scales. The inclusion of non-native interactions is important, as residual frustration can play an integral role in protein dynamics [167].

The effective energy parameters, ϵ and δ (see Equation 4.1), for a given amino-acid sequence are designed in such a way that the protein is able to fold to its native state. The parameters $\epsilon(c_i, c_j)$ and $\delta(c_i, c_j)$ for native SH3 are generated through an iterative procedure, described in detail in ref [85] and Chapter 4. The optimal set of parameters $\{\epsilon_{opt}, \delta_{opt}\}$ is defined by the maximum energy gap criterion [85, 86, 87] on a large set of decoys (i.e. low energy, compact, non-native structures) such that $\Delta E(\{\epsilon_{opt}\}, \{\delta_{opt}\}) = \max_{\{\epsilon\},\{\delta\}} \left[\left(\min_{i\in(1,\ldots,N_{decoys})} E_{\Gamma_i}\right) - E_{\Gamma_{nat}}\right]$, where $E_{\Gamma_{nat}}$ refers to the energy of the crystal structure Γ_{nat} of native SH3 and E_{Γ_i} is the energy corresponding to the decoy structure Γ_i. Any low energy configuration with RMSD > 2.5 Å from the crystal structure is considered a decoy structure. An initial training set of 1000 decoys is used and a total of 60 iterations is performed. Multiple short "heat-and-quench" unfolding/refolding simulations are carried out to check if the obtained set of parameters is able to refold the protein to the correct native state (within RMSD < 2.5 Å from the crystal structure of native SH3).

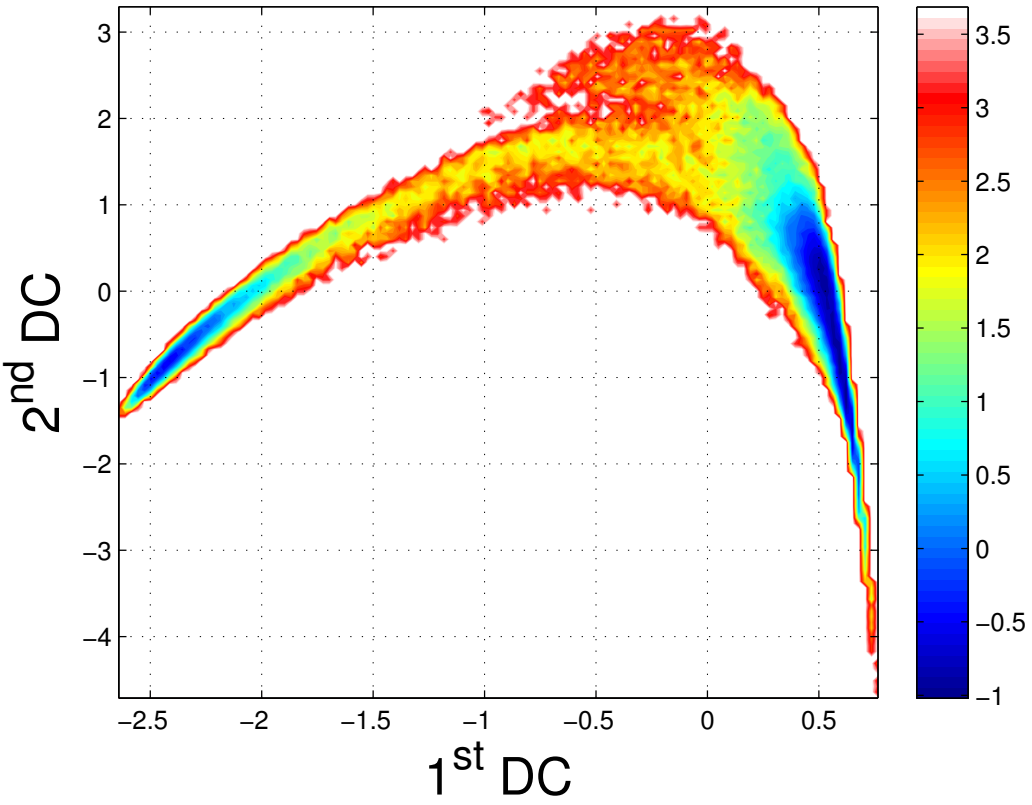


Figure 5.2 : Free energy landscape for the vacuum DMC model using the 1^{st} and 2^{nd} diffusion coordinates determined by LSDMap. Colors are measured in units of RT. The folded state populates the left basin and the unfolded state populates the right basin.

5.2.2 Incorporating Experimental Data

When an optimal set of parameters is obtained that is able to consistently fold the protein, the parameters are further optimized to match existing experimental data, as described in refs [170, 106]. An extensive sampling of the the free energy landscape is analyzed with the Locally Scaled Diffusion Map (LSDMap) procedure [95, 96], described in detail in Chapter 4.

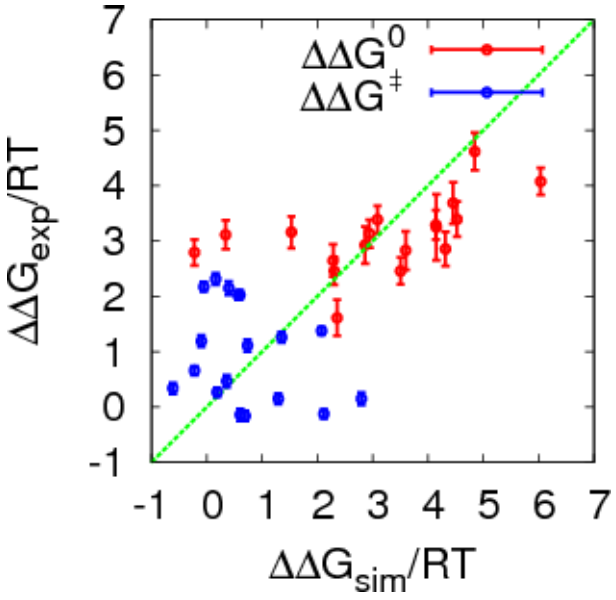


Figure 5.3 : Comparison of experimentally measured [168, 169] free energy differences of mutation with the computationally measured values from the vacuum DMC model after only 3 iterations.

The native (N), unfolded (U), and transition (T) ensembles are defined as the basins and barrier, respectively, of the free energy landscape in terms of the first and second diffusion coordinates (DC) of LSDMap. For each ensemble, the effect of mutating a single residue is estimated using Free Energy Perturbation Theory assuming that a mutation results in a minimal disturbance of the landscape. To ease the description of the mutations, the Hamiltonian H (see Equation 4.1), is split in to repulsive and attractive parts, *Rep* and *Att*, respectively:

$$Rep = 5\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} \tag{5.1}$$

$$Att = -6\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10} \tag{5.2}$$

$$H = \epsilon(c_i, c_j)\left[Rep + \delta(c_i, c_j)Att\right] \tag{5.3}$$

The interaction parameters are perturbed by a conservative mutation in which the size of the new bead is less than or equal to the size of the mutated bead. A mutation involves a recalculation of the Hamiltonian with different interaction parameters for the chosen bead, resulting in Rep' and Att'. The perturbation to the Hamiltonian is then:

$$\delta H = \epsilon(c_i, c_j)\left[(Rep' - Rep) + \delta(c_i, c_j)(Att' - Att)\right] \tag{5.4}$$

The ΔG of the mutation is calculated from the ensemble average of the Hamiltonian perturbation above as:

$$\Delta G = -RT \ln \left\langle e^{\frac{-\delta H}{RT}} \right\rangle \tag{5.5}$$

and $\Delta\Delta G^0$ and $\Delta\Delta G^{\ddagger}$ are calculated as the difference between the unfolded ensemble average with the native and transition ensemble averages, respectively:

$$\Delta\Delta G^0 = \Delta G^N - \Delta G^U \tag{5.6}$$

$$\Delta\Delta G^{\ddagger} = \Delta G^T - \Delta G^U \tag{5.7}$$

Based on this perturbation, corrections are made to the interaction parameters to match the experimentally measured values for $\Delta\Delta G^0$ and $\Delta\Delta G^{\ddagger}$ [168, 169]:

$$\frac{\delta H}{\delta \epsilon} = [(Rep' - Rep) + \delta(c_i, c_j)(Att' - Att)] \tag{5.8}$$

$$\frac{\delta \Delta G}{\delta \epsilon} = \frac{\left\langle [(Rep' - Rep) + \delta(c_i, c_j)(Att' - Att)]\, e^{\frac{-\delta H}{RT}} \right\rangle}{\left\langle e^{\frac{-\delta H}{RT}} \right\rangle} \tag{5.9}$$

$$\Delta\Delta G_{exp} - \Delta\Delta G_{sim} = \frac{\delta \Delta\Delta G}{\delta \epsilon} \Delta\epsilon \tag{5.10}$$

After a few iterations the correlation of the experimentally and computationally measured free energy differences is improved considerably (see Figure 5.3). The resulting parameters are then used as an initial guess to define a solvated CG protein model.

5.2.3 Extending the Coarse Graining to the Solvent

A representative native structure is generated by energy minimization of the all-atom pdb structure in SPC/E water at 300K. Only the C_α sites of the all-atom protein and the Oxygen sites of the three-site SPC/E model are retained, resulting in roughly 9000 solvent molecules as a 21^{st} color type surrounding the coarse grain protein. This coarse grained representation of the all-atom energy minimized system is used as the new native target for the interaction parameter optimization. Because the interaction of the solvent with itself is fixed (the solvent model is the same as the one designed for the adaptive resolution study of hydrophobic solutes [38], described in Chapter 3), additional parameters to describe the interactions of the protein with the solvent are all that is necessary for the new optimization.

Solvated decoys are added to the training set developed for the vacuum system, and the optimization is performed on both the protein-protein and protein-solvent interactions simultaneously. After 50 iterations, more than 200,000 decoys were collected to maximize the energy gap. The resulting parameter sets were then used to explore the landscape at a folding temperature of 335K. For each parameter set a total of 50 different simulations of 200 ns (enough for more than one folding/unfolding transition in each simulation) were combined and analyzed with LSDMap to generate diffusion coordinates.

The first and second diffusion coordinates from the LSDMap analysis were again used to define folded, unfolded, and transition ensembles. The interaction parameters are corrected in the same fashion as the vacuum model, and after 3 iterations the fit is acceptable (see Figure 5.5).

5.3 Results and Conclusion

The vacuum model of SH3 has been studied before [85, 95], and the vacuum results from the present study do not deviate from these previous results (see Figure 5.2). The unfolded state is extremely flexible with no clear structure retained.

The solvated model differs from the vacuum model in several key areas (see Figure 5.4). The unfolded state is more collapsed than in the vacuum model, allowing for local structure to be temporarily formed. The folded state (A) does not exhibit water in the core, as expected. During the unfolding process, the RT loop opens a

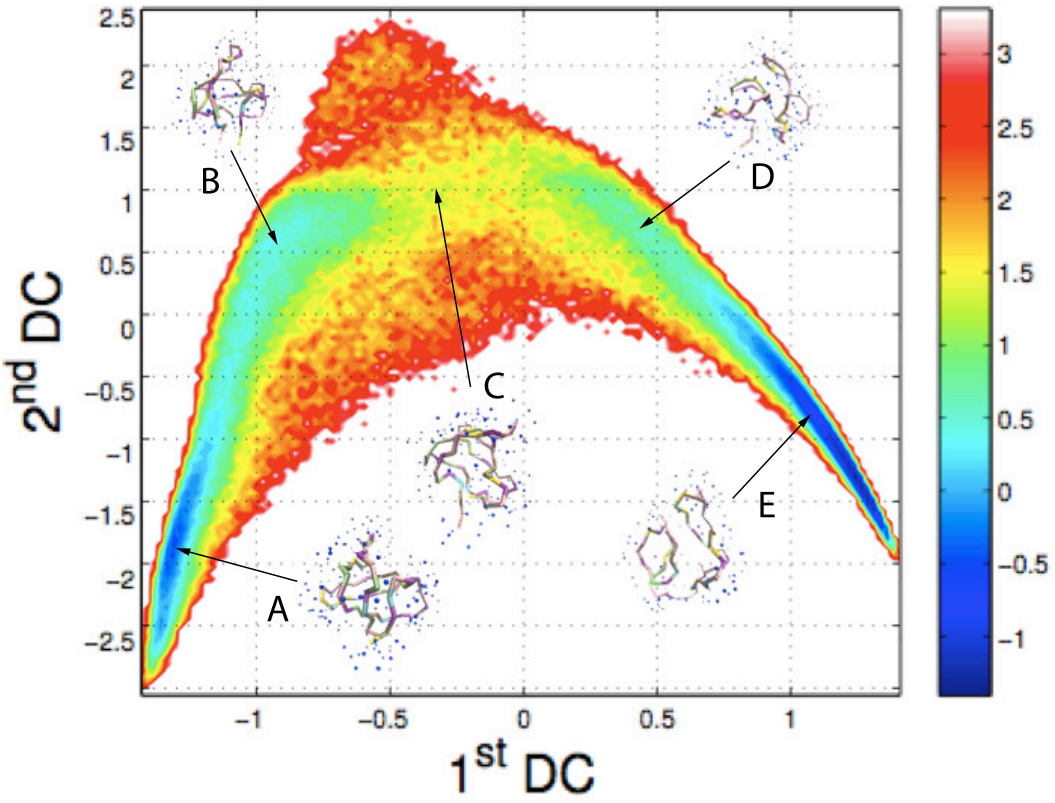


Figure 5.4 : Free energy landscape for the solvated DMC model using the 1^{st} and 2^{nd} diffusion coordinates determined by LSDMap. Colors are measured in units of RT. The folding/unfolding pathway is labelled with representative structures: A) the folded configuration, no water in core; B) the RT loop opening to fill with 5-10 water molecules; C) the beta sheets separating, orthogonality being distorted; D) the beta sheets mediated by 10-20 water molecules; E) the beta sheets randomly organized, structure forms and breaks rapidly.

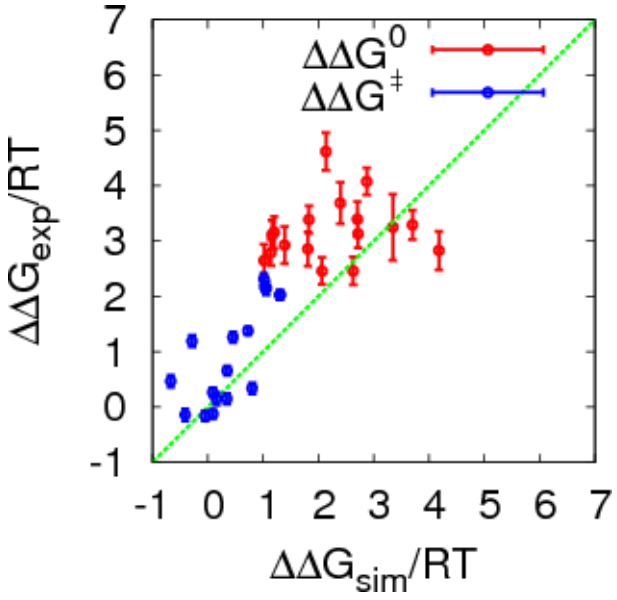


Figure 5.5 : Comparison of experimentally measured [168, 169] free energy differences of mutation with the computationally measured values from the solvated DMC model after only 3 iterations.

pocket (B) which allows water to begin to fill. As the two beta sheets are separated (C), a transition is achieved only when the two can be sufficiently separated to allow complete disruption of their orthogonal orientation (D). The unfolded state consists of collapsed random arrangements of these two beta sheet loops (E).

The folding process follows a similar mechanism as the unfolding process. Simulations started from fully elongated states (not shown) immediately collapse to a water mediated collapsed structure (E) and local search is made for the beta sheet connection. Once the sheets have arranged themselves orthogonally (B-C), water is quickly expulsed and the protein finally collapses to the most compact native structure (A). Comparing the free energy landscapes of the vacuum and solvated models

(Figures 5.2 and 5.4), it is clear that only one narrow pathway is dominant for the vacuum model, while the solvated model exhibits at least two folding intermediates along a broad pathway. Multiple folding intermediates have also been reported for SH3 domains in earlier work with all-atom and structure-based models [171, 172, 173] as well as experimental studies with relaxation dispersion NMR [127, 174], amyloid studies [126], and high denaturant unfolding [128, 175].

Because the vacuum model is free to move without hindrance from the environment, the transition state in the vacuum model only approximately resembles the collapsed unfolded structures of the solvated model. The critical step of the folding/unfolding process is the expulsion of water, which cannot be captured by the vacuum model without modifying the potential to include a desolvation barrier [129].

We have shown the importance of explicit water for modeling biological systems. Without an aqueous environment to support the protein, unfolded structures do not explore the transition state in a water mediated way. When solvent is explicitly included, the folding/unfolding process can be accurately observed and representative transition states can be determined. The observed folding steps for SH3 involve an organization of beta sheets and a water expulsion step, allowing for multiple folding intermediates. The combined protein-solvent model results encourage the further application of adaptive resolution techniques to resolve important regions of proteins in all-atom detail while using a coarser description in the rest of the system to conserve computational resources.

Chapter 6

Summary

This work has accomplished many of the preliminary steps which are necessary to define an adaptive multi-scale model for solvated biomolecular systems. It was shown that the AdResS methodology could be extended to systems with hydrophobic molecules, and that it is possible to couple a coarse grained model for proteins with a single-site model for water. Reconstruction methods were demonstrated in the use of coarse grain models to provide starting points for targeted search with all-atom models.

The final piece of the adaptive biological model will be the coupling of an all-atom model of a protein with a coarse grain model like the one described in this thesis. Early work in this area has been recently reported for a solvated polymer [176]. Any hybrid model of a protein could be readily integrated in to an AdResS setup which resolves a portion of the protein and it's local environment at the all-atom level while coarse graining away faster degrees of freedom far from the site of interest, allowing for microsecond sampling of the free energy landscape.

Bibliography

[1] D. Huang and D. Chandler, "Temperature and length scale dependence of hydrophobic effects and their possible implications for protein folding," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 97, no. 15, p. 8324, 2000.

[2] D. Chandler, "Interfaces and the driving force of hydrophobic assembly," *Nature*, vol. 437, no. 7059, p. 640, 2005.

[3] K. Lum, D. Chandler, and J. Weeks, "Hydrophobicity at small and large length scales," *J. Phys. Chem. B*, vol. 103, p. 4570, 1999.

[4] K. Dill, "Theory for the folding and stability of globular proteins," *Biochemistry*, vol. 24, p. 1501, 1985.

[5] V. Petrenko, M. Dubova, Y. Kessler, and M. Perova, "Water in computer experiment: Contradiction in parameterization of potentials," *Russ. J. Phys. Chem.*, vol. 74, p. 1777, 2000.

[6] J. Bernal and R. Fowler, "A theory of water and ionic solution, with particular reference to hydrogen and hydroxyl ions," *J. Chem. Phys.*, vol. 1, p. 515, 1933.

[7] A. Rahman and F. Stillinger, "Molecular dynamics study of liquid water," *J. Chem. Phys.*, vol. 55, no. 7, p. 3336, 1971.

[8] W. Jorgensen, J. Chandrasekhar, J. Madura, R. Impey, and M. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, no. 2, p. 926, 1983.

[9] H. Berendsen, J. Grigera, and T. Straatsma, "The missing term in effective pair potentials," *J. Phys. Chem.*, vol. 91, p. 6269, 1987.

[10] J. Zielkiewicz, "Structural properties of water: Comparison of the spc, spce, tip4p, and tip5p models of water," *J. Chem. Phys.*, vol. 123, p. 104501, 2005.

[11] W. Jorgensen and J. Tirado-Rives, "Potential energy functions for atomic-level simulations of water and organic and biomolecular systems," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, no. 19, p. 6665, 2005.

[12] W. Still, A. Tempczyk, R. Hawley, and T. Hendrickson, "Semianalytical treatment of solvation for molecular mechanics and dynamics," *J. Am. Chem. Soc.*, vol. 112, no. 16, p. 6127, 1990.

[13] C. Tan, Y. Tan, and R. Luo, "Implicit nonpolar solvation models," *J. Chem. Phys. B*, vol. 111, p. 12263, 2007.

[14] S. Matysiak and C. Clementi, "Optimal combination of theory and experiment for the characterization of the protein folding landscape of s6: how far can a minimalist model go?," *J. Mol. Biol.*, vol. 343, p. 235, 2004.

[15] S. Izvekov, M. Parrinello, C. Burnham, and G. Voth, "Effective force fields for

condensed phase systems from ab initio molecular dynamics simulation: A new method for force-matching," *J. Chem. Phys.*, vol. 120, p. 10896, 2004.

[16] S. Izvekov and G. Voth, "Multiscale coarse graining of liquid-state systems," *J. Chem. Phys.*, vol. 123, p. 134105, 2005.

[17] S. Izvekov, J. Swanson, and G. Voth, "Coarse-graining in interaction space: A systematic approach for replacing long-range electrostatics with short-range potentials," *J. Phys. Chem. B*, vol. 112, no. 15, p. 4711, 2008.

[18] D. Reith, M. Puetz, and F. Mueller-Plathe, "Deriving effective mesoscale potentials from atomistic simulations," *J. Comput. Chem. Phys.*, vol. 24, p. 1624, 2003.

[19] H. Wang, C. Junghans, and K. Kremer, "Comparative atomistic and coarse-grained study of water: What do we lose by coarse-graining?," *Eur. Phys. J. E*, vol. 28, no. 2, p. 221, 2009.

[20] A. Lyubartsev and A. Laaksonen, "Calculation of effective interaction potentials from radial distribution functions: A reverse monte carlo approach," *Phys. Rev. E*, vol. 52, no. 4, p. 52, 1995.

[21] S. Matysiak, C. Clementi, M. Praprotnik, K. Kremer, and L. Delle Site, "Modeling diffusive dynamics in adaptive resolution simulation of liquid water," *J. Chem. Phys.*, vol. 128, p. 024503, 2008.

[22] M. Praprotnik, L. Delle Site, and K. Kremer, "Adaptive resolution molecular-dynamics simulation: Changing the degrees of freedom on the fly," *J. Chem. Phys.*, vol. 123, p. 224106, 2005.

[23] B. Ensing, S. Nielsen, P. Moore, M. Klein, and M. Parrinello, "Energy conservation in adaptive hybrid atomistic/coarse-grain molecular dynamics," *J. Chem. Theory Comput.*, vol. 3, p. 1100, 2007.

[24] L. Delle Site, "Some fundamental problems for an energy conserving adaptive resolution molecular dynamics scheme," *Phys. Rev. E*, vol. 76, p. 047701, 2007.

[25] M. Praprotnik, K. Kremer, and L. Delle Site, "Adaptive molecular resolution via a continuous change of the phase space dimensionality," *Phys. Rev. E*, vol. 75, p. 017701, 2007.

[26] M. Praprotnik, L. Delle Site, and K. Kremer, "Adaptive resolution scheme for efficient hybrid atomistic-mesoscale molecular dynamics simulations of dense liquids," *Phys. Rev. E*, vol. 73, p. 066701, 2006.

[27] M. Praprotnik, L. Delle Site, and K. Kremer, "A macromolecule in a solvent: Adaptive resolution molecular dynamics simulation," *J. Chem. Phys.*, vol. 126, p. 134902, 2007.

[28] S. Matysiak, M. Praprotnik, L. Delle Site, K. Kremer, and C. Clementi, "Adaptive resolution simulation of liquid water," *J. Phys. Condens. Matter*, vol. 19, p. 292201, 2007.

[29] M. Neumann, "Dipole moment fluctuation formulas in computer simulations of polar systems," *Mol. Phys.*, vol. 50, no. 4, p. 841, 1983.

[30] M. Neumann, "The dielectric constant of water. computer simulations with the mcy potential," *J. Chem. Phys.*, vol. 82, no. 12, p. 5663, 1985.

[31] G. Papoian, J. Ulander, M. Eastwood, Z. Luthey-Schulten, and P. Wolynes, "Water in protein structure prediction," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 101, no. 10, pp. 3352–3357, 2004.

[32] P. Ball, "Water as an active constituent in cell biology," *Chem. Rev.*, vol. 108, pp. 74–108, 2008.

[33] C. Clementi, "Coarse-grained models of protein folding: Toy-models or predictive tools?," *Curr. Opin. Struct. Biol.*, vol. 18, no. 1, pp. 10–15, 2008.

[34] T. Lazaridis and M. Paulaitis, "Entropy of hydrophobic hydration: a new statistical mechanical formulation," *J. Phys. Chem.*, vol. 96, p. 3847, 1992.

[35] C. Lee, J. McCammon, and P. Rossky, "The structure of liquid water at an extended hydrophobic surface," *J. Chem. Phys.*, vol. 80, pp. 4448–4455, 1984.

[36] A. Patel, P. Varilly, and D. Chandler, "Fluctuations of water near extended hydrophobic and hydrophilic surfaces," *J. Phys. Chem. B*, vol. 114, p. 1954, 2010.

[37] S. Poblete, M. Praprotnik, K. Kremer, and L. Delle Site, "Coupling different levels of resolution in molecular simulations," *J. Chem. Phys.*, vol. 132, no. 11, p. 114101, 2010.

[38] B. Lambeth, C. Junghans, K. Kremer, C. Clementi, and L. Delle Site, "Communication: On the locality of hydrogen bond networks at hydrophobic interfaces," *J. Chem. Phys.*, vol. 133, p. 221101, 2010.

[39] S. Sarupria and S. Garde, "Quantifying water density fluctuations and compressibility of hydration shells of hydrophobic solutes and proteins," *Phys. Rev. Lett.*, vol. 103, p. 037803, 2009.

[40] N. Choudhury and B. Pettitt, "The dewetting transition and the hydrophobic effect," *J. Am. Chem. Soc.*, vol. 129, pp. 4847–4852, 2007.

[41] H. Ashbaugh and M. Paulaitis, "Effect of solute size and solute-water attractive interactions on hydration water structure around hydrophobic solutes," *J. Am. Chem. Soc.*, vol. 123, pp. 10721–10728, 2001.

[42] J. Ma, D. Alfe, A. Michaelides, and E. Wang, "The water-benzene interaction: Insight from electronic structure theories," *J. Chem. Phys.*, vol. 130, p. 154303, 2009.

[43] D. Chandler, J. Weeks, and H. Andersen, "Van der waals picture of liquids, solids, and phase transformations," *Science*, vol. 220, pp. 787–794, 1983.

[44] D. Weiss, T. Raschke, and M. Levitt, "How hydrophobic buckminsterfullerene affects surrounding water," *J. Phys. Chem. B*, vol. 112, p. 2981, 2008.

[45] C. Xu and G. Scuseria, "An o(n) tight-binding study of carbon clusters up to c_{8640}: the geometrical shape of the giant icosahedral fullerenes," *Chem. Phys. Lett.*, vol. 262, pp. 219–226, 1996.

[46] R. Zope, T. Baruah, M. Pederson, and B. Dunlap, "Static dielectric response of icosahedral fullerenes from c_{60} to c_{2160} characterized by an all-electron density functional theory," *Phys. Rev. B*, vol. 77, p. 115452, 2008.

[47] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation," *J. Chem. Theory Comput.*, vol. 4, no. 3, p. 435, 2008.

[48] B. Lambeth, C. Junghans, K. Kremer, and C. Clementi, "To be decided," *J. Chem. Theory Comput.*, vol. to be submitted, 2010.

[49] V. Rühle, C. Junghans, A. Lukyanov, K. Kremer, and D. Andrienko, "Versatile object-oriented toolkit for coarse-graining applications," *J. Chem. Theory Comput.*, vol. 5, pp. 3211–3223, 2009.

[50] R. Lynden-Bell and T. Head-Gordon, "Solvation in modified water models: toward understanding hydrophobic solvation," *Mol. Phys.*, vol. 104, pp. 3593–3605, 2006.

[51] J. Errington and P. Debenedetti, "Relationship between structural order and the anomalies of liquid water," *Nature*, vol. 409, no. 6818, p. 318, 2001.

[52] L. Scatena, M. Brown, and G. Richmond, "Water at hydrophobic surfaces: Weak hydrogen bonding and strong orientation effects," *Science*, vol. 292, pp. 908–912, 2001.

[53] J. Mittal and G. Hummer, "Static and dynamic correlations in water at hydrophobic interfaces," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 105, no. 51, pp. 20130–20135, 2008.

[54] A. Rahman and F. Stillinger, "Hydrogen-bond patterns in liquid water," *J. Am. Chem. Soc.*, vol. 95, no. 24, p. 7943, 1973.

[55] N. Choudhury and B. Pettitt, "Local density profiles are coupled to solute size and attractive potential for nanoscopic hydrophobic solutes," *Mol. Simul.*, vol. 31, no. 6-7, pp. 457–463, 2005.

[56] V. R. Rao and D. D. Oprian, "Activating mutations of rhodopsin and other g protein-coupled receptors," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 25, pp. 287–314, 1996.

[57] K. G. Bache, T. Slagsvold, and H. Stenmark, "Defective downregulation of receptor tyrosine kinases in cancer," *EMBO J*, vol. 23, pp. 2707–2712, 2004.

[58] W. W. Sprenger, W. D. Hoff, J. P. Armitage, and K. J. Hellingwerf, "The eu-

bacterium *Ectothiorhodospira halophila* is negatively phototactic, with a wavelength dependence that fits the absorption spectrum of the photoactive yellow protein," *J. Bacteriol.*, vol. 175, pp. 3096–3104, 1993.

[59] G. E. O. Borgstahl, D. R. Williams, and E. D. Getzoff, "1.4 å structure of photoactive yellow protein, a cytosolic photoreceptor - unusual fold, active-site, and chromophore," *Biochemistry*, vol. 34, pp. 6278–6287, 1995.

[60] M. Baca, G. E. O. Borgstahl, M. Boissinot, P. M. Burke, D. R. Williams, K. A. Slater, and E. D. Getzoff, "Complete chemical structure of photoactive yellow protein: Novel thioester-linked 4-hydroxycinnamyl chromophore and photocycle chemistry," *Biochemistry*, vol. 33, pp. 14369–14377, 1994.

[61] R. Kort, H. Vonk, X. Xu, W. D. Hoff, W. Crielaard, and K. J. Hellingwerf, "Evidence for the trans-cis isomerization of the p-coumaric acid chromophore as the photochemical basis of the photocycle of photoactive yellow protein," *FEBS Lett.*, vol. 382, pp. 73–78, 1996.

[62] G. Groenhof, M. Bouxin-Cademartory, B. Hess, S. P. de Visser, H. J. C. Berendsen, M. Olivucci, A. E. Mark, and M. A. Robb, "Photoactivation of the photoactive yellow protein: Why photon absorption triggers a trans-to-cis isomerization of the chromophore in the protein," *J. Am. Chem. Soc.*, vol. 126, pp. 4228–4233, 2004.

[63] W. D. Hoff, A. Xie, I. H. M. van Stokkum, X. J. Tang, J. Gural, A. R. Kroon,

and K. J. Hellingwerf, "Global conformational changes upon receptor stimulation in photoactive yellow protein," *Biochemistry*, vol. 38, pp. 1009–1017, 1999.

[64] Y. Imamoto, H. Kamikubo, M. Harigai, N. Shimizu, and M. Kataoka, "Light-induced global conformational change of photoactive yellow protein in solution," *Biochemistry*, vol. 41, pp. 1009–1017, 2002.

[65] B. C. Lee, A. Pandit, P. A. Croonquist, and W. D. Hoff, "Folding and signaling share the same pathway in a photoreceptor," *Proc. Natl. Acad. Sci. USA*, vol. 98, pp. 9062–9067, 2001.

[66] B. Lee, P. A. Croonquist, T. R. Sosnick, and W. D. Hoff, "PAS domain receptor photoactive yellow protein is converted to a molten globule state upon activation," *J. Biol. Chem.*, vol. 276, pp. 20821–20823, 2001.

[67] C. Bernard, K. Hoube, N. M. Derix, D. Marks, M. A. van der Horst, K. J. Hellingwerf, R. Boelans, R. Kaptein, and A. J. van Nuland, N, "The solution structure of a transient photoreceptor intermediate: Delta25 photoactive yellow protein," *Structure*, vol. 13, pp. 953–962, 2005.

[68] G. Rubinstenn, G. W. Vuister, F. A. A. Mulder, P. E. Dux, R. Boelens, K. J. Hellingwerf, and R. Kaptein, "Structural and dynamic changes of photoactive yellow protein during its photocycle in solution," *Nature Struct. Biol.*, vol. 5, pp. 568–570, 1998.

[69] B. Lee, P. A. Croonquist, and W. D. Hoff, "Mimic of photocycle by a protein folding reaction in photoactive yellow protein," *J. Biol. Chem.*, vol. 276, pp. 44481–44487, 2001.

[70] H. Ihee, S. Rajagopal, V. Srajer, R. Pahl, S. Anderson, M. Schmidt, F. Schotte, P. A. Anfinrud, M. Wulff, and K. Moffat, "Visualizing reaction pathways in photoactive yellow protein from nanoseconds to seconds," *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 7145–7150, 2005.

[71] S. Anderson, V. Srajer, R. Pahl, S. Rajagopal, F. Schotte, P. Anfinrud, M. Wulff, and K. Moffat, "Chromophore conformation and the evolution of tertiary structural changes in photoactive yellow protein," *Structure*, vol. 12, pp. 1039–1045, 2004.

[72] U. K. Genick, S. M. Soltis, P. Kuhn, I. L. Canestrelli, and E. D. Getzoff, "Structure at 0.85 å resolution of an early protein photocycle intermediate," *Nature*, vol. 392, pp. 206–209, 1998.

[73] U. K. Genick, G. E. O. Borgstahl, K. Ng, Z. Ren, C. Pradervand, P. M. Burke, V. Srajer, T. Teng, W. Schildkamp, D. E. McRee, K. Moffat, and E. D. Getzoff, "Structure of a protein photocycle intermediate by millisecond time-resolved crystallography," *Science*, vol. 275, pp. 1471–1474, 1997.

[74] G. Groenhof, M. F. Lensink, J. G. Snijders, H. J. C. Berendsen, and A. E. Mark, "Signal transduction in the photoactive yellow protein. i. photon absorption and

the isomerization of the chromophore," *Proteins: Struct. Funct. Genet.*, vol. 48, pp. 202–211, 2002.

[75] G. Groenhof, M. F. Lensink, H. J. C. Berendsen, and A. E. Mark, "Signal transduction in the photoactive yellow protein. ii. proton transfer initiates conformational changes," *Proteins: Struct. Funct. Genet.*, vol. 48, pp. 212–219, 2002.

[76] G. Groenhof, M. Bouxin-Cademartory, B. Hess, S. P. de Visser, H. J. C. Berendsen, M. Olivucci, A. E. Mark, and M. A. Robb, "Photoactivation of the photoactive yellow protein: Why photon absorption triggers a trans-to-cis isomerization of the chromophore in the protein," *J. Am. Chem. Soc.*, vol. 126, pp. 4228–4233, 2004.

[77] I. Antes, W. Thiel, and van Gunsteren W. F., "Molecular dynamics simulations of photoactive yellow protein (pyp) in three states of its photocycle: A comparison with x-ray and nmr data and analysis of the effects of glu46 deprotonation and mutation," *Euro. Biophys. J.*, vol. 31, pp. 504–520, 2002.

[78] S. Arai, M. Togashi, M. Shiozawa, Y. Inoue, and M. Sakurai, "Molecular dynamics simulation of the m intermediate of photoactive yellow protein in the crystalline state," *Chem. Phys. Lett*, vol. 414, pp. 230–233, 2005.

[79] M. Shiozawa, M. Yoda, N. Kamiya, N. Asakawa, J. Higo, Y. Inoue, and M. Sakurai, "Evidence for large structural fluctuations of the photobleached interme-

diate of photoactive yellow protein in solution," *J. Am. Chem. Soc.*, vol. 123, pp. 7445–7446, 2001.

[80] J. Vreede, W. Crielaard, K. J. Hellingwerf, and P. G. Bolhuis, "Predicting the signaling state of photoactive yellow protein," *Biophys. J.*, vol. 88, pp. 3525–3535, 2005.

[81] J. Vreede, J. Juraszek, and P. Bolhuis, "Predicting the reaction coordinates of millisecond light-induced conformational changes in photoactive yellow protein," *Proc. Natl. Acad. Sci. USA*, vol. 107, pp. 2397–2402, 2010.

[82] N. Gō, "Theoretical studies of protein folding," *Annu. Rev Biophys. and Bioeng.*, vol. 12, pp. 183–210, 1983.

[83] C. Clementi, H. Nymeyer, and J. N. Onuchic, "Topological and energetic factors: What determines the structural details of the transition state ensemble and en-route intermediates for protein folding? An investigation for small globular proteins," *J. Mol. Biol.*, vol. 298, pp. 937–953, 2000.

[84] P. Das, C. Wilson, G. Fossati, P. Wittung-Stafshede, K. Matthews, and C. Clementi, "Characterization of the folding landscape of monomeric lactose repressor: Quantitative comparison of theory and experiment," *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 14569–14574, 2005.

[85] P. Das, S. Matysiak, and C. Clementi, "Balancing energy and entropy: A minimalist model for the characterization of protein folding landscapes," *Proc. Natl.*

Acad. Sci. USA, vol. 102, pp. 10141–10146, 2005.

[86] C. Clementi, A. Maritan, and J. R. Banavar, "Folding, design, and determination of interaction potentials using off-lattice dynamics of model heteropolymers," *Phys. Rev. Lett.*, vol. 81, pp. 3287–3290, 1998.

[87] R. Goldstein, Z. Luthey-Schulten, and P. Wolynes, "Optimal protein-folding codes from spin-glass theory," *Proc. Natl. Acad. Sci. USA*, vol. 82, pp. 4918–4922, 1992.

[88] A. Xie, W. D. Hoff, A. R. Kroon, and K. J. Hellingwerf, "Glu46 donates a proton to the 4-hydroxycinnamate anion chromophore during the photocycle of photoactive yellow protein," *Biochemistry*, vol. 35, pp. 14671–14678, 1996.

[89] A. Xie, L. Kelemen, J. Hendriks, B. J. White, K. J. Hellingwerf, and W. D. Hoff, "Formation of a new buried charge drives a large-amplitude protein quake in photoreceptor activation," *Biochemistry*, vol. 40, pp. 1510–1517, 2001.

[90] R. Brudler, T. E. Meyer, U. K. Genick, S. Devanathan, T. T. Woo, D. P. Millar, K. Gerwert, M. A. Cusanovich, G. Tollin, and E. D. Getzoff, "Coupling of hydrogen bonding to chromophore conformation and function in photoactive yellow protein," *Biochemistry*, vol. 39, pp. 13478–13486, 2000.

[91] U. K. Genick, S. Devanathan, T. E. Meyer, I. L. Canestrelli, E. Williams, M. A. Cusanovich, G. Tollin, and E. D. Getzoff, "Active site mutants impli-

cate residues for control of color and light cycle kinetics of photoactive yellow protein," *Biochemistry*, vol. 36, pp. 8–14, 1997.

[92] K. Itoh and M. Sasai, "Dynamical transitions and proteinquake in photoactive yellow protein," *Proc. Natl. Acad. Sci. USA*, vol. 101, pp. 14736–14741, 2004.

[93] A. M. Ferrenberg and R. H. Swendsen, "Optimized monte carlo data analysis," *Phys. Rev. Lett.*, vol. 63, pp. 1185–1198, 1989.

[94] A. M. Ferrenberg and R. H. Swendsen, "New monte carlo technique for studying phase transitions," *Phys. Rev. Lett.*, vol. 61, pp. 2635–2638, 1988.

[95] M. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, "Determination of reaction coordinates via locally scaled diffusion map," *J. Chem. Phys.*, vol. 134, no. 12, p. 124116, 2011.

[96] W. Zheng, M. Rohrdanz, M. Maggioni, and C. Clementi, "Polymer reversal rate calculated via locally scaled diffusion map," *J. Chem. Phys.*, vol. 134, no. 14, p. 144109, 2011.

[97] K. J. Hellingwerf, J. Hendriks, and T. Gensch, "Photoactive yellow protein, a new type of photoreceptor protein: Will this "yellow lab" bring us where we want to go?," *J. Phys. Chem.*, vol. 107, pp. 1082–1094, 2003.

[98] C. J. Craven, N. M. Derix, J. Hendricks, R. Boelens, K. J. Hellingwerf, and R. Kaptein, "Probing the nature of the blue-shifted intermediate of photoactive

yellow protein in solution by nmr: Hydrogen-deuterium exchange data and ph studies," *Biochemistry*, vol. 39, pp. 14392–14399, 2000.

[99] G. Rubinstenn, G. W. Vuister, F. A. A. Mulder, P. E. Dux, K. J. Boelens, R. Hellingwerf, and R. Kaptein, "Structural and dynamic changes of photoactive yellow protein during its photocycle in solution," *Nature Struct. Biol.*, vol. 5, pp. 568–570, 1998.

[100] J. Hendricks, T. Gensch, M. A. van der Horst, K. J. Hellingwerf, and J. J. van Thor, "Transient exposure of hydrophobic surface in the photoactive yellow protein monitored with Nile red," *Biophys. J.*, vol. 82, pp. 1632–1643, 2002.

[101] J. Hendriks, W. D. Hoff, W. Crielaard, and K. J. Hellingwerf, "Protonation/deprotonation reactions triggered by photoactivation of photoactive yellow protein from *Ectothiorhodospira halophila*," *J. Biol. Chem.*, vol. 274, pp. 17655–17660, 1999.

[102] J. S. Khan, Y. Imamoto, M. Kataoka, F. Tokunaga, and M. Terazima, "Time-resolved thermodynamics: Heat capacity change of the transient species during photoreaction of pyp," *J. Am. Chem. Soc.*, vol. 128, pp. 1002–1008, 2006.

[103] J. Hendriks, T. Gensch, L. Hviid, M. A. van der Horst, K. J. Hellingwerf, and J. J. van Thor, "Transient exposure of hydrophobic surface in the photoactive yellow protein monitored with nile red," *Biophys. J.*, vol. 82, pp. 1632–1643, 2002.

[104] N. M. Derix, R. W. Wechselberger, M. A. van der Horst, K. J. Hellingworf, R. Boelens, R. Kaptein, and N. A. J. van Nuland, "Lack of negative charge in the E46Q mutant of photoactive yellow protein prevents partial unfolding of the blue-shifted intermediate," *Biochemistry*, vol. 42, pp. 14501–14506, 2003.

[105] G. Fuentes, A. J. Nederveen, R. Kaptein, R. Boelens, and A. M. Bonvin, "Describing partially unfolded states of proteins from sparse nmr data.," *J. Biomol. NMR*, vol. 33, pp. 175–186, 2005.

[106] S. Matysiak and C. Clementi, "Minimalist protein model as a diagnostic tool for misfolding and aggregation," *J. Mol. Biol.*, vol. 363, pp. 297–308, 2006.

[107] D. E. Otzen, O. Kristensen, and M. Oliveberg, "Designed protein tetramer zipped together with a hydrophobic alzheimer homology: A structural clue to amyloid assembly," *Proc. Natl. Acad. Sci. USA*, vol. 97, pp. 9907–9912, 2000.

[108] G. Groenhof, M. F. Lensink, H. J. C. Berendsen, and A. E. Mark, "Signal transduction in the photoactive yellow protein. ii. proton transfer initiates conformational changes," *Proteins: Struct. Funct. Genet.*, vol. 48, pp. 212–219, 2002.

[109] E. Leenders, L. Guidoni, J. Vreede, U. Rothlisberger, P. Bolhuis, and E. Meijer, "Protonation of the chromophore in the photoactive yellow protein," *J. Phys. Chem. B*, vol. 111, pp. 3765–3774, 2007.

[110] J. Vreede, K. Hellingwerf, and P. Bolhuis, "Helix formation is a dynamical bottleneck in the recovery reaction of photoactive yellow protein," *Proteins: Struct. Funct. Genet.*, vol. 72, pp. 136–149, 2008.

[111] M. Praprotnik, L. Delle Site, and K. Kremer, "Adaptive resolution molecular-dynamics simulation: changing the degrees of freedom on the fly," *J. Chem. Phys.*, vol. 123, no. 22, p. 224106, 2005.

[112] M. Neri, C. Anselmi, M. Cascella, A. Maritan, and P. Carloni, "Coarse-grained model of proteins incorporating atomistic detail of the active site," *Phys Rev Lett*, vol. 95, pp. 218102–218102, Nov 2005.

[113] Q. Shi, S. Izvekov, and G. A. Voth, "Mixed atomistic and coarse-grained molecular dynamics: Simulation of a membrane-bound ion channel," *J. Phys. Chem. B*, vol. 110, pp. 15045–15048, 2006.

[114] E. Lyman, F. M. Ytreberg, and D. M. Zuckerman, "Resolution exchange simulation," *Phys. Rev. Lett.*, vol. 96, pp. 028105–028105, Jan 2006.

[115] M. Christen and W. F. van Gunsteren, "Multigraining: an algorithm for simultaneous fine-grained and coarse-grained simulation of molecular systems," *J. Chem. Phys.*, vol. 124, p. 154106, Apr 2006.

[116] W. Kwak and U. H. Hansmann, "Efficient sampling of protein structures by model hopping," *Phys. Rev. Lett.*, vol. 95, no. 13, p. 138102, 2005.

[117] Z. Fan, J. Hwang, and A. Warshel, "Using simplified protein representation as a reference potential for all-atom calculations of folding free energy," *Theor. Chem. Acc.*, vol. 103, pp. 77–80, 1999.

[118] G. De Mori, G. Colombo, and C. Micheletti, "Study of the villin headpiece folding dynamics by combining coarse-grained monte carlo evolution and all-atom molecular dynamics," *Proteins*, vol. 58, pp. 459–471, Feb 2005.

[119] M. Rossi, V. Blum, P. Kupser, G. von Helden, F. Bierau, K. Pagel, G. Meijer, and M. Scheffler, "Secondary structure of Ac-Ala$_n$-LysH$^+$ polyalanine peptides (n=5,10,15) in vacuum: Helical or not?," *J. Phys. Chem. Lett.*, vol. 1, pp. 3465–3470, 2010.

[120] E. Cocinero, P. Carcabal, T. Vaden, J. Simons, and B. Davis, "Sensing the anomeric effect in a solvent-free environment," *Nature*, vol. 469, pp. 76–80, 2011.

[121] B. Hess, S. Len, N. van der Vegt, and K. Kremer, "Long time atomistic polymer trajectories from coarse grained simulations: bisphenol-a polycarbonate," *Soft Matter*, vol. 2, no. 5, pp. 409–414, 2006.

[122] C. Peter and K. Kremer, "Multiscale simulations of soft matter systems - from the atomistic to the coarse-grained level and back," *Soft Matter*, vol. 5, pp. 4357–4366, 2009.

[123] A. Heath, L. Kavraki, and C. Clementi, "From coarse-grain to all-atom: Toward multiscale analysis of protein landscapes," *Proteins*, vol. 68, no. 3, pp. 646–661, 2007.

[124] R. L. Dunbrack, "Rotamer libraries in the 21st century," *Curr. Opin. Struct. Biol.*, vol. 12, no. 4, pp. 431–440, 2002.

[125] P. Ledbetter and C. Clementi, "A new perspective on transition states: chi1 separatrix," *J. Chem. Phys.*, vol. 135, no. 4, p. 044116, 2011.

[126] S. Ventura, J. Zurdo, S. Narayanan, M. Parreño, R. Mangues, B. Reif, F. Chiti, E. Giannoni, C. Dobson, F. Aviles, and L. Serrano, "Short amino acid stretches can mediate amyloid formation in globular proteins: the src homology 3 (sh3) case," *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 19, pp. 7258–63, 2004.

[127] D. Korzhnev, X. Salvatella, M. Vendruscolo, A. Di Nardo, A. Davidson, C. Dobson, and L. Kay, "Low-populated folding intermediates of fyn sh3 characterized by relaxation dispersion nmr," *Nature*, vol. 430, no. 6999, pp. 586–590, 2004.

[128] A. Wani and J. Udgaonkar, "Revealing a concealed intermediate that forms after the rate-limiting step of refolding of the sh3 domain of pi3 kinase," *J. Mol. Biol.*, vol. 387, no. 2, pp. 348–62, 2009.

[129] M. S. Cheung, A. E. Garcia, and J. N. Onuchic, "Protein folding mediated by solvation: Water expulsion and formation of the hydrophobic core occur after the structural collapse," *Proc. Natl. Acad. Sci. USA*, vol. 99, pp. 685–690, 2002.

[130] R. Paliwal and E. London, "Comparison of the conformation, hydrophobicity, and model membrane interactions of diphtheria toxin to those of formaldehyde-treated toxin (diphtheria toxoid): Formaldehyde stabilization of the native conformation inhibits changes that allow membrane insertion," *Biochemistry*, vol. 35, pp. 2374–2379, 1996.

[131] B. Vecsey-Semjen, R. Mollby, and F. G. van der Goot, "Partial c-terminal unfolding is required for channel formation by staphylococcal alpha-toxin," *J. Biol. Chem.*, vol. 271, pp. 8655–8660, 1996.

[132] S. M. Harper, J. M. Christie, and K. H. Gardner, "Disruption of the lov-j alpha helix interaction activates phototropin kinase activity," *Biochemistry*, vol. 43, pp. 16184–16192, 2004.

[133] W. D. Hoff, I. H. M. van Stokkum, J. Gural, and K. J. Hellingwerf, "Comparison of acid denaturation and light activation in the eubacterial blue-light receptor photoactive yellow protein," *Biochim. Biophys. Acta*, vol. 1332, pp. 151–162, 1997.

[134] V. Thammavongsa, L. Mancino, and M. Raghavan, "Polypeptide substrate recognition by calnexin requires specific conformations of the calnexin protein," *J. Biol. Chem.*, vol. 39, pp. 33497–33505, 2005.

[135] B. K. Das and J. J. Liang, "Detection and characterization of alpha-crystallin intermediate with maximal chaperone-like activity," *Biochem. Biophys. Res.*

Commun., vol. 236, pp. 370–374, 1997.

[136] A. Shehu, C. Clementi, and L. E. Kavraki, "Modeling protein conformational ensembles: From missing loops to equilibrium fluctuations," *Proteins: Struct. Funct. Genet.*, vol. 65, pp. 164–179, 2006.

[137] A. Shehu, L. E. Kavraki, and C. Clementi, "On the characterization of protein native state ensembles," *Biophys. J.*, vol. 92, pp. 1503–1511, 2007.

[138] A. Shehu, C. Clementi, and L. E. Kavraki, "Computing protein conformations from a single structure: Modeling protein flexibility at equilibrium," *Algorithmica*, vol. 48, pp. 303–327, 2007.

[139] M. Praprotnik, L. Delle Site, K. Kremer, S. Matysiak, and C. Clementi, "Adaptive resolution simulation of liquid water," *J. Phys. Condens. Matter.*, vol. 19, p. 292201, 2007.

[140] A. P. Heath, L. E. Kavraki, and C. Clementi, "From coarse-grain to all-atom: Toward multiscale analysis of protein landscapes," *Proteins: Struct. Funct. Genet.*, vol. 68, pp. 646–661, 2007.

[141] L. Maragliano and E. Vanden-Eijnden, "A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations," *Chem. Phys. Lett.*, vol. 426, pp. 168–175, 2006.

[142] C. Abrams and E. Vanden-Eijnden, "Large-scale conformational sampling of

proteins using temperature-accelerated molecular dynamics," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 107, pp. 4961–4966, 2010.

[143] I. G. Kevrekidis, C. W. Gear, and G. Hummer, "Equation-free: The computer-aided analysis of complex multiscale systems," *AIChE J.*, vol. 50, 2004.

[144] P. Doruker, R. L. Jernigan, and I. Bahar, "Dynamics of large proteins through hierarchical levels of coarse-grained structures," *J. Comput. Chem.*, vol. 23, pp. 119–127, 2002.

[145] V. Tozzini, J. Trylska, C. Chang, and J. A. McCammon, "Flap opening dynamics in HIV-1 protease explored with a coarse-grained model," *J. Struct. Bio.*, vol. 157, pp. 606–615, 2007.

[146] F. Tama and C. L. Brooks, "Symmetry, form, and shape: Guiding principles for robustness in macromolecular machines," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 35, pp. 115–133, 2006.

[147] S. Izvekov and G. A. Voth, "A multiscale coarse-graining method for biomolecular systems," *J. Phys. Chem. B*, vol. 109, pp. 2469–2473, 2005.

[148] A. Liwo, M. Khalili, and H. A. Scheraga, "Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptied chains," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, pp. 2362–2367, 2005.

[149] V. Molinero and W. A. Goddard III, "M3B: A coarse grain force field for molecular simulations of malto-oligosaccharides and their water mixtures," *J. Phys. Chem. B*, vol. 108, pp. 1414–1427, 2004.

[150] W. Han and Y. Wu, "Coarse-grained protein model coupled with a coarse-grained water model: Molecular dynamics study of polyalanine-based peptides," *J. Chem. Theory Comput.*, vol. 3, pp. 2146–2161, 2007.

[151] W. Shinoda, R. DeVane, and M. L. Klein, "Multi-property fitting and parameterization of a coarse grained model for aqueous surfactants," *Mol. Simul.*, vol. 33, pp. 27–36, 2007.

[152] R. DeVane, W. Shinoda, P. B. Moore, and M. L. Klein, "Transferable coarse grain nonbonded interaction model for amino acids," *J. Chem. Theory Comput.*, vol. 5, pp. 2115–2124, 2009.

[153] X. He, W. Shinoda, R. DeVane, K. L. Anderson, and M. L. Klein, "Paramaterization of a coarse-grained model for linear alkylbenzene sulfonate surfactants and molecular dynamics studies of their self-assembly in aqueous solution," *Chem. Phys. Lett.*, vol. 487, pp. 71–76, 2010.

[154] N. Basdevant, D. Borgis, and T. HaDuong, "A coarse-grained protein-protein potential derived from an all-atom force field," *J. Phys. Chem. B*, vol. 111, pp. 9390–9399, 2007.

[155] T. HaDuong, N. Basdevant, and D. Borgis, "A polarizable coarse-grained water model for coarse-grained proteins simulations," *Chem. Phys. Lett.*, vol. 468, pp. 79–82, 2009.

[156] A. Rzepiela, L. Schafer, N. Goga, H. J. Risselada, A. DeVries, and S. Marrink, "Reconstruction of atomistic details from coarse-grained structures," *J. Comput. Chem.*, vol. 31, pp. 1333–1343, 2010.

[157] A. J. Rzepiela, D. Sengupta, N. Goga, and S. J. Marrink, "Membrane poration by antimicrobial peptides combining atomistic and coarse-grained descriptions," *Faraday Discuss.*, vol. 144, pp. 431–443, 2010.

[158] K. Hadley and C. McCabe, "On the investigation of coarse-grained models for water: Balancing computational efficiency and the retention of structural properties," *J. Phys. Chem. B*, vol. 114, pp. 4590–4599, 2010.

[159] J. C. Shelley, M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay, and M. L. Klein, "A coarse grain model for phospholipid simulation," *J. Phys. Chem. B*, vol. 105, pp. 4464–4470, 2001.

[160] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S. J. Marrink, "The MARTINI coarse grained forcefield: extension to proteins," *J. Chem. Theory Comput.*, vol. 4, pp. 819–834, 2008.

[161] A. Chandra and T. Ichiye, "Dynamical properties of the soft sticky dipole model

of water: Molecular dynamics simulations," *J. Chem. Phys.*, vol. 111, pp. 2701–2709, 1999.

[162] V. Molinero and E. B. Moore, "Water modeled as an intermediate element between carbon and silicon," *J. Phys. Chem. B*, vol. 113, pp. 4008–4016, 2009.

[163] J. E. Shea, J. N. Onuchic, and C. L. Brooks, "Probing the folding free energy landscape of the src-SH3 protein domain," *Proc. Natl. Acad. Sci. USA*, vol. 99, pp. 16064–16068, 2002.

[164] C. Clementi, "Coarse-grained models of protein folding: Toy-models or predictive tools?," *Curr. Opin. Struct. Biol.*, vol. 18, pp. 10–15, 2008.

[165] C. Clementi, A. E. García, and J. N. Onuchic, "Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: A all-atom representation study of protein l," *J. Mol. Biol.*, vol. 326, pp. 933–954, 2003.

[166] Y. Levy, S. S. Cho, J. N. Onuchic, and P. G. Wolynes, "A survey of flexible protein binding mechanisms and their transition states using native topology based energy landscapes," *J. Mol. Biol.*, vol. 346, pp. 1121–1145, 2005.

[167] D. U. Ferreiro, J. A. Hegler, E. A. Komives, and P. G. Wolynes, "Localizing frustration in native proteins and protein assemblies," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 104, pp. 19819–19824, 2007.

[168] V. Grantcharova, D. Riddle, J. Santiago, and D. Baker, "Important role of hydrogen bonds in the structurally polarized transition state for folding of the src sh3 domain," *Nature Struct. Biol.*, vol. 5, no. 8, pp. 714–20, 1998.

[169] D. Riddle, V. Grantcharova, J. Santiago, E. Alm, I. Ruczinski, and D. Baker, "Experiment and theory highlight role of native state topology in sh3 folding," *Nature Struct. Biol.*, vol. 6, no. 11, pp. 1016–24, 1999.

[170] S. Matysiak and C. Clementi, "Optimal combination of theory and experiment for the characterization of the protein folding landscape of s6: How far can a minimalist model go," *J. Mol. Biol.*, vol. 343, pp. 235–248, 2004.

[171] J. M. Borreguero, F. Ding, S. V. Buldyrev, H. E. Stanley, and N. V. Dokholyan, "Multiple folding pathways of the sh3 domain," *Biophys. J.*, vol. 87, no. 1, pp. 521–533, 2004.

[172] F. Ding, W. Guo, N. V. Dokholyan, E. I. Shakhnovich, and J.-E. Shea, "Reconstruction of the src-sh3 protein domain transition state ensemble using multiscale molecular dynamics simulations," *J. Mol. Biol.*, vol. 350, no. 5, pp. 1035–1050, 2005.

[173] a. R. Lam, J. M. Borreguero, F. Ding, N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, and E. Shakhnovich, "Parallel folding pathways in the sh3 domain protein," *J. Mol. Biol.*, vol. 373, no. 5, pp. 1348–1360, 2007.

[174] D. Korzhnev, P. Neudecker, A. Zarrine-Afsar, A. Davidson, and L. Kay, "Abp1p and fyn sh3 domains fold through similar low-populated intermediate states," *Biochemistry*, vol. 45, no. 34, pp. 10175–10183, 2006.

[175] A. H. Wani and J. B. Udgaonkar, "Native state dynamics drive the unfolding of the sh3 domain of pi3 kinase at high denaturant concentration," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 106, no. 49, pp. 20711–20716, 2009.

[176] M. Praprotnik, S. Poblete, and K. Kremer, "Statistical physics problems in adaptive resolution computer simulations of complex fluids," *J. Stat. Phys.*, vol. 145, pp. 946–966, 2011.